IDENTIFICATION AND ANALYSIS OF IMPRINTED AND NON-IMPRINTED GENES IN DISTAL HUMAN CHROMOSOME 20q13

Verónica Fabiola Morán Barroso

Degree of Doctor of Philosophy
University of Edinburgh
2001

DECLARATION

I declare that:

a) This thesis has been composed by myself.

b) The work is either my own or the work/author involved is clearly stated.

ACKNOWLEDGMENTS

I want to express my gratitude to my supervisor, Professor David T. Bonthron, who guided me through my studies and who with kindness and patience read the drafts of this thesis.

I want also to express my thankfulness to the Mexican National Council for Science and Technology, CONACYT, for awarding me the scholarship that allowed me to carry out my studies.

During these years, I have been lucky to make good friends in Edinburgh and in Leeds, for your support and friendship, I will always be grateful to all of you. In particular I want to thank to Dr. Bruce E. Hayward and Dr. Lisa Strain for their help, and support.

I want to thank my parents and my brothers, for their support, despite being thousands of miles away, they walked this path with me.

I thank my friends back at home in Mexico, for their support especially Dr. Susana Kofman-Alfaro for her motivation and encouragement for many years.

I am also grateful to my family and friends at home who once a year pretended (rather convincingly, I must say) to be very happy and pleased to see me as a sort of "new year comet" for a few days.

ABSTRACT

It is now well-established that certain regions in the human genome show allelic non-equivalence, depending on their maternal or paternal origin. This phenomenon, known as genomic imprinting, violates the classical Mendelian assumption of equivalent parental genetic contributions at autosomal loci. Imprinting occurs through the action of epigenetic mechanisms during gametogenesis, which lead to parent-of-origin specific transcriptional regulation of a subset of genes in the offspring.

This thesis describes work that arose from studies of the imprinting of *GNAS1*, on human chromosome 20q13. Null mutations in *GNAS1* cause the hormone-resistance syndrome pseudohypoparathyroidism type 1a (PHP1a). It was demonstrated that *GNAS1* is imprinted, as predicted from the anomalous inheritance of PHP1a, but that its allele-specific regulation is highly complex. This gene is shown to encode several protein products: (i) the alpha subunit of the stimulatory G protein G_s; this protein is biallelically derived; (ii) NESP55, a neuroendocrine secretory protein, expressed exclusively from the maternal allele; (iii) XLαs, a Golgi-specific G protein that is expressed exclusively from the paternal allele.

Many known imprinted genes lie in clusters that may span several hundred of kb, and may be co-ordinately regulated by an imprinting control region. To assess whether *GNAS1* is part of such an imprinted gene cluster, genomic clones were analysed for the presence of other nearby transcripts. This resulted in the identification of two novel genes, *CTSZ* and *CGI107*, as well as a number of putative transcripts that lie within *GNAS1*. One of the latter forms part of a spliced polyadenylated antisense transcript that spans the upstream region of *NESP55*.

The *CTSZ* gene was shown to comprise 6 exons and 5 introns, spanning ~12 kb. It encodes cathepsin Z, a cysteine protease whose precise function is undefined. The *CGI107* gene was shown to comprise 6 exons and 5 introns, spanning ~10 kb, encoding a 194 amino acid protein of unknown function, but showing sequence similarity to the *kisir* protein in *Drosophila*. Four chromosomally dispersed processed pseudogenes of *CGI107* were identified.

Expressed sequence polymorphisms were identified within both genes and used to assess the allelic origin of transcripts in a range of human tissues. Biallelic expression was demonstrated for both genes, making it unlikely that they are imprinted. Together with the biallelic expression of another neighbouring gene, *TH1*, these results suggest the possibility that *GNAS1* may not after all be part of an extensive cluster of imprinted genes. This has implications for further studies of the mechanism of imprinting control in this region.

ABBREVIATIONS

Aa	Amino acid
ADP	Adenosine 5′-diphosphate
AHO	Albright hereditary osteodystrophy
AMP	Adenosine 5′-monophosphate
AS	Angelman syndrome
AS-SRO	Shortest region of deletion overlap in AS
ATP	Adenosine 3′, 5′- cyclic phosphate ("cyclic AMP)
cAMP	Cyclic adenosine 5′-monophosphate
bp	Base pair
BWS	Beckwith-Wiedemann syndrome
BWSCR	BWS breakpoint cluster regions
CTX	Cholera toxin
DMRs	Differentially methylated regions
DR	Direct repeat
EDTA	Ethylenediamine tetraacetic acid
GAPs	GTPase-activating proteins
GEFs	Guanine nucleotide exchange factors
GDP	Guanosine 5′-diphosphate
GPCRs	G protein-coupled receptors
$G_s\alpha$	α-subunit of the heterotrimeric stimulatory G protein
gsp	G stimulatory oncogene
GTP	Guanosine 5′-triphosphate
HATs	Histone acetyltransferases
HDACs	Histone deacetylases
HP	Hypoparathyroidism
IC	Imprinting centre
IPTG	Isopropylthio-β-D˙galactoside
kb	kilobases
LH	Luteinizing hormone
LOH	Loss of heterozygosity

LOI	Loss of imprinting
MARs	Matrix attachment regions
MAS	McCune-Albright syndrome
Mb	Megabase
MI	Meiosis I
MII	Meiosis II
MW	Molecular weight
NAD	Nicotamide adenine dinucleotide
ND	Nondisjunction
NDGA	Nordihydroguaiaretic acid
nt	Nucleotide
PDE	Phosphodiesterase
PFGE	Pulsed-field electrophoresis gel
PGCs	Female primordial germ cells
PHP	Pseudohypoparathyroidism
PLA_2	Phospholipase A_2
PLC	Phospholipase C
PTH	Parathyroid hormone
PTHrP	PTH-related peptide
PTH1R	PTH type 1 receptor
PTX	Pertussis toxin
PWS	Prader-Willi syndrome
PWS-SRO	Shortest region of overlap of microdeletions in PWS
RGS	Regulators of G protein signalling
RLGS-M	Restriction landmark genomic scanning for methylation differences
SDS	Sodium dodecyl sulphate
TS	Tumour suppressor genes
TSH	Thyroid-stimulating hormone
UPD	Uniparental disomy
UTR	Untranslated region
VNTR	Variable number of tandem repeats
XIC	X inactivation centre

LIST OF EXPERIMENTS.

CHAPTER 3. *GNAS1*, AN IMPRINTED GENE IN HUMAN CHROMOSOME 20q13

EXPERIMENT	CARRIED OUT BY
Identification of the A20 exon by RLGS-M	Dr. Y. Hayashizaki (Japan)
Analysis of A20 sequence as part of *GNAS1*	Professor Bonthron
Sequencing of cDNA clone 827-k06	V. Moran
Screening of the RPCI1 library	V. Moran
Construction of subclone libraries dJ309f20	V.Moran
Sequencing subclone *Sac*I-*Asc*I Number 4	Dr. B. Hayward.
Sequencing of subclone *Bam*HI 21	V. Moran
Localisation and sequencing of exon 3N in *GNAS1*	V. Moran
Differential methylation analysis in the XLαs and NESP55 regions of *GNAS1*	Dr. Lisa Strain.
Monoallelic expression analysis of RT-PCR products	Dr. B Hayward
Identification and analysis of *GNAS1* EST sequences in the databases	Dr. B. Hayward
5′ RACE analysis on fetal muscle RNA and identification of the NESP55 sequence	Dr. B. Hayward
PFGE with *Mlu*I-*Not*I digested fragments and hybridisation of NESP55 and Xlαs probes	Dr. B. Hayward
Restriction map of subclone Bgl34	V. Moran
Sequencing analysis of Bgl34	V. Moran

CHAPTER 4. SCREENING FOR NEW GENES IN SUBCLONE LIBRARIES DERIVED FROM PAC DNA CLONES THAT MAP TO CHROMOSOME 20q13

EXPERIMENT	CARRIED OUT BY
Analysis screening of *Bam*HI/*Bgl*II library including BamA1, G11, C6, F3 and BglC2, F1 and D7	V. Moran
Analysis of cDNA clone 1430	V. Moran
Analysis of clone 088f21 (latter identified as containing exons IV and V of the antisense transcript).	V. Moran
Analysis of clone 1476 k11	V. Moran
Analysis and construction of dJ309f20 subclone library with vector TVEC	V. Moran
Analysis of subclone Hin4F	V.Moran
Subcloning of *Not*I and *Xba*I fragments into the vector pCRII	Dr. B. Hayward.
Identification of subclone 309M13-3 (SP6-end)	Dr. B. Hayward
PCR screening of RCPI1 PAC library for subclones mapping at the SP6 end of dj309f20 and hibridisation screening of high	V. Moran

density gridded filters for RCPI1 library	
Construction and analysis of the dJ96n2 and dJ654c22 Pac libraries	V. Moran

CHAPTER 5 (CATHEPSIN Z (*CTSZ*), A NON-IMPRINTED GENE DOWNSTREAM OF *GNAS1* IN 20q13) AND CHAPTER 6 (*CGI-107*, A NON-IMPRINTED GENE DOWNSTREAM OF *GNAS1* IN 20q13).

Preparation of cDNA for RT-PCR analysis, some were already available, the rest were prepared by Dr. B. Hayward. The rest of the experiments were carried out by myself.

CONTENTS **Page No.**

CHAPTER 2. MATERIAL AND METHODS

CHAPTER 3. *GNAS1*, AN IMPRINTED GENE IN HUMAN CHROMOSOME 20q13

CHAPTER 5. CATHEPSIN Z (*CTSZ*), A NON-IMPRINTED GENE DOWNSTREAM OF *GNAS1* IN 20q13

CHAPTER 1

THE GENOMIC IMPRINTING MECHANISM

CHAPTER 1. THE GENOMIC IMPRINTING MECHANISM

1.1 INTRODUCTION

Based on observations of his famous breeding experiments in the plant *Pisum sativum*, Gregor Mendel, in 1865, concluded the existence of several rules that govern the patterns by which genes are passed from one generation to the next. One of the cornerstones of Mendel's "laws" is the idea that each parent makes a similar contribution of genetic material to their offspring (Hall, 1990). In accordance with this, most genes in humans are inherited in two copies, one from each parent (the clear exceptions being sex-linked genes in males and mitochondrial genes) (Nicholls, 2000).

However, since the 1980s, evidence has emerged that cuts across these Mendelian principles, since it has been found that certain regions in the genome are functionally non-equivalent, depending on their maternal or paternal allelic origin. This phenomenon, known as genomic imprinting, is believed to depend on epigenetic modifications of the genome. Imprinting may be defined as a gametic process, which leads to a parent-of-origin specific transcriptional silencing of a subset of genes in the offspring (Barlow, 1995, Greally, 1998, Caspary et al., 1998). It has been estimated that there are 100-200 imprinted genes in the mammalian genome (Barlow, 1995).

Genomic imprinting is considered to be an *epigenetic* process in the sense that two alleles that may be identical in DNA sequence are regulated differently within the same nucleus. Also, the process is entirely reversible, the silent allele becoming reactivated when passed through the germ line of the opposite parental sex and conversely the active allele becoming silenced (Feil and Kelsey, 1997).

Empirical evidence that reflects the existence of genomic imprinting was gathered in ancient times, probably more than 3000 years ago, among the mule breeders of Asia Minor (Savory, 1970, Morison and Reeve, 1998). It is still a familiar fact that there are phenotypic differences in the offspring obtained from crossing horses and donkeys, depending on the sex of the parents. A cross between a male donkey and a

female horse produces a mule, whereas the offspring of the reciprocal cross, between a female donkey and a male horse, is known as a hinny (Savory, 1970). Fertility studies have led to the general conclusion that the male hinny and mule are both sterile. In contrast, fertile cases of female mule and hinny have both been documented (Rong et al., 1988). It is thought that the structural and numerical differences between the parental chromosomal sets (2n=64 or 62, in horse and donkey respectively) result in abnormal pairing during male meiosis, and that the resulting meiotic failure underlies the sterility observed (Chandley et al., 1974). This example conforms to the Haldane Rule regarding interspecific hybrids, that states that the heterogametic sex is the one more likely to be sterile.

A formal demonstration of unequal genetic contributions from parents to their offspring came first from the observations made by Dawson (1965) on the outcomes of interspecific crosses between *Peromyscus maniculatus* (the deer mouse) and *Peromyscus polionotus* (the oldfield mouse). It was observed that the F1 hybrid from the female *P. maniculatus*/male *P. polionotus* cross was significantly smaller on average than *P. polionotus*, the smaller parental form. On the other hand, the F1 hybrid from the reciprocal cross was generally larger than either parent, demonstrating a clear non-equivalence in the parental influence on the size of offspring. Recent reinvestigation of this phenomenon has resulted in the identification of both genetic and epigenetic factors that contribute to the *Peromyscus* hybrid phenotypes. These include alterations in X chromosome inactivation and loss of imprinting of maternally and paternally imprinted genes in both species (Vrana et al., 2000).

The term parental "imprinting" was first used in this genetic context in reference to whole chromosomes, originally to refer to the selective elimination of paternal chromosomes in the insect *Schiara* (Crouse, 1960). Later, this term was used to describe the selective inactivation of paternally derived X chromosomes in extraembryonic membranes in the mouse (Lyon and Rastan, 1984).

During the 1970s and 1980s, several different lines of experimental and clinical evidence were accumulated to support the existence of genomic imprinting in mammals. These may be grouped as follows (Hall 1997):

i) Pronuclear transplantation and parthenogenesis experiments in mice
ii) Differing phenotypes of triploids in humans
iii) Expression of chromosomal uniparental disomies in humans and mice
iv) Expression of chromosomal deletions in humans
v) Parental origin-specific expression of trangenes in mice
vi) Expression of specific genes in humans and mice, particularly related to growth, development and behaviour

Examples of each type of evidence are summarized in the following sections.

1.1.1 PRONUCLEAR TRANSPLANTATION AND PARTHENOGENESIS EXPERIMENTS IN MICE

In the 1980's, several series of pronuclear transplantation experiments were carried out, with the aim of investigating differences between the roles played by the maternal and paternal genomic contributions to the offspring. McGrath and Solter (1984a) transplanted single pronuclei between one-cell-stage mouse embryos. Removal of one pronucleus, followed by the introduction of a second from another embryo, was used to construct biparental diploid gynogenetic, androgenetic and control (one male and one female pronucleus) embryos. It was found that both gynogenetic and androgenetic embryos were unable to complete normal embryogenesis, whereas control embryos sometimes did complete normal development. This indicated a requirement for both a maternal and a paternal genetic contribution to allow normal development, at least in the mouse.

Similar experiments were carried out by Surani et al. (1984). These workers showed that while 40-50% of heterozygous reconstituted eggs with a male and a female pronucleus developed to term, none of the eggs with two male pronuclei did so.

Embryos in the latter case were very retarded, although the trophoblast developed relatively well, compared with embryos having two female pronuclei. It was concluded that the paternal genome was essential for the normal development of the extraembryonic tissues, while the maternal genome was essential for embryogenesis (Barton et al., 1984, Mann and Lovell-Badge, 1984).

McGrath and Solter (1984b) performed further pronuclear transplants between mouse embryos carrying the T^{hp} mutation in the T-associated maternal effect (*Tme*) locus on chromosome 17. The effects of this mutation are determined by the sex of the parent from which it is inherited; when inherited from the female parent it is lethal, but embryos that inherit it from the male parent survive. The authors noted that the maternal-lethal effect persisted when T^{hp}/+ pronuclei were transplanted into +/+ cytoplasm, suggesting that the epigenetic defect underlying the anomalous inheritance pattern was localized to the nucleus. They therefore suggested that a specific gene(s) on chromosome 17 is inherited in a functional form from one parent but in a non-functional form from the other. Later on, Barlow et al. (1991) demonstrated that the imprinted gene *Igf2r* was indeed linked to the *Tme* locus.

Naturally-occurring human equivalents of the pronuclear transplantation experiments in mice have been studied. These are two proliferative disorders that result from uniparental conceptions, ovarian teratomas and complete hydatidiform moles. Teratomas are embryonic tumours that contain tissues from all three embryonic germ layers but completely lack placental tissue. Ovarian teratomas have been demonstrated to have two maternally-derived haploid chromosome sets, in other words to be parthenogenetic (Linder et al., 1975, Parrington et al., 1984).

In complete hydatiform moles, in contrast, no embryo is formed and there is an abundance of abnormal placental tissue. These complete moles are almost always androgenetic in origin, the most frequent mechanism probably being the fertilisation of an oocyte that lacks a maternal pronucleus, followed by a duplication of the paternal chromosomes (Kajii and Ohama, 1977, Devriendt, 2000).

1.1.2 PHENOTYPES OF TRIPLOIDS IN HUMANS

Somewhat less extreme imprinting effects are also observed in the phenotypes of triploid human embryos. These differ depending on the paternal origin of the extra haploid set of chromosomes, a fact first recognised by McFadden and Kalousek (1991) and McFadden et al. (1993). These authors used DNA polymorphisms to show whether the triploidy in 9 human fetuses was maternal or paternal in origin. They described two phenotypes. The type I phenotype was characterised by a well-grown fetus with proportionate head size or slight microcephaly; the placenta was large and had the appearance of a partial hydatiform mole. In these type I cases, the extra set of chromosomes was paternal. In contrast, the type II phenotype showed a very growth retarded fetus with relative macrocephaly; the placentas were very small and did not have molar characteristics. In these cases, the extra set of chromosomes was maternal. No significant differences were observed between the two groups regarding other malformations seen in triploidy, such as syndactyly and neural tube defects (McFadden et al., 1993).

1.1.3 EXPRESSION OF CHROMOSOMAL UNIPARENTAL DISOMIES IN HUMANS AND MICE

Uniparental disomy (or UPD) refers to the circumstance in which both of a pair of chromosomal homologues are inherited from one parent, the total number of chromosomes being normal (Hall, 1997). The origin of UPD individuals depends mostly on nondisjunction (ND) events, that can occur either in meiosis I (MI) or meiosis II (MII). There are three subsequent likely routes that lead to UPD (Soler et al., 2000):

i) *Trisomic zygote rescue*- the origin of a trisomic fetus followed by the loss of one chromosome.

ii) *Gametic complementation*- the fertilisation of a disomic egg by a nullisomic sperm or *vice versa*.

iii) *Postzygotic duplication* – compensatory UPD, a somatic event leading to replacement of an abnormal or absent chromosome with the normal.

In mice, UPD can be experimentally induced. Cattanach and Kirk (1985) intercrossed mice that were heterozygous for robertsonian translocations of chromosomes 2, 11, and 13, such that UPD was the only viable outcome of specific crosses. They showed that the offspring had different phenotypes depending on the chromosomal disomy involved. Mice with disomy 13 were normal, but in contrast the maternal duplication/paternal deficiency for chromosome 2 was lethal. Disomy 11 mice showed differences in size: when their chromosomes 11 were of maternal origin they were consistently smaller and when these chromosomes were of paternal origin they were consistently larger than normal.

These experiments show that the differential functioning of the maternal and paternal genomes results from the combined effects of parental non-equivalence for several chromosomal regions. They also show that imprinting affects gene activity in a selective but reproducible way. Finally, since there were examples of normal development among the uniparental disomies studied, it seemed that some chromosomes were probably devoid of imprinted genes with important developmental functions (Cattanach and Kirk, 1985, Cattanach, 1986).

In humans, UPD has been observed for most of the chromosomes, and the pathological effects of some of these are well-defined. The most notable example is UPD of chromosome 15, which results in two distinct clinical disorders, depending on its parental origin: Prader Willi syndrome (PWS) results when both chromosome 15s are maternal (matUPD), while conversely patUPD15 results in Angelman syndrome (AS) (Ledbetter et al., 1981, Nicholls et al., 1989, Mascar et al., 1992, Smith et al., 1997).

The consequences of UPD do not necessarily result from imprinting effects, since isodisomy (in which two copies of a single parental chromosomal homologue are inherited) also conveys the risk of unmasking an autosomal recessive disease, if the

chromosome in question carries an abnormal allele (Hall, 1997). However, an imprinting effect may be safely assumed if three or more cases of UPD produce a similar phenotype, not attributable to other known causes (recessive trait or mosaicism). On the other hand, even a single case of UPD with a normal phenotype constitutes relatively strong evidence against a major imprinting effect for that specific chromosome (Ledbetter and Engel, 1995).

1.1.4 PARENTAL ORIGIN-SPECIFIC EXPRESSION OF TRANSGENES IN MICE

The first opportunity to study individual imprinted loci arose from the observation of parental origin-specific expression of some mouse transgenes (Swain et al., 1987, Sapienza et al., 1987, Reik et al., 1987). Swain et al. (1987) produced a transgenic mouse in which an activated *c-myc* transgene was autosomally inherited but showed parent-specific expression. When inherited from the male parent, the transgene was expressed in the heart but no other tissue; on the other hand, when maternally inherited, the same transgene was not expressed at all. Furthermore, the imprinted pattern of expression also correlated with a parent-of-origin specific methylation state (the maternal allele being methylated).

Sapienza et al. (1987) and Reik et al. (1987) further demonstrated that the methylation pattern of transgenes could be changed by switching their gamete of origin in successive generations. This is a requirement that must be satisfied if methylation is postulated to play the key role in genomic imprinting, of establishing the parent-specific epigenetic mark; see section 1.1.8.1.

1.1.5 DIFFERENTIAL EXPRESSION OF CHROMOSOMAL DELETION IN HUMANS

Once again, PWS and AS represent the best-defined examples of differential expression of chromosome deletions in humans. Almost 70% of individuals with PWS have a deletion of chromosome 15q11-13, and it is always the paternal

chromosome that is deleted. On the other hand, about 70% of AS individuals have a similar del(15)(q11-13), but in this case it is always maternally derived. These observations complement those on whole-chromosome UPD15 (see above). They demonstrate that a common region (at least at the Mb resolution of the deletion analysis) is involved in both disorders, and that this critical region functions differently on the paternal and the maternal chromosome 15 (Ledbetter et al., 1981, Knoll et al., 1989, Cassidy, 1997, Hall, 1997, Christian et al., 1998).

1.1.6 EXPRESSION OF SPECIFIC IMPRINTED GENES IN HUMANS AND MICE

There are several examples of clinical disorders that appear to be inherited more frequently from one parent than from the other, or with a difference in phenotype, suggesting the possibility that genomic imprinting affects the chromosome regions (and genes) involved in these conditions. Some of these loci are listed in Table 1.1 (Mannens et al., 1994, Morison and Reeves, 1998, Morison et al., 2001).

In addition, it should be mentioned that there are other conditions (such as fragile X syndrome and Huntington disease) in which a parent-of-origin effect is apparent on transmission of the disease allele. Both these disorders result from trinucleotide repeat expansions, though their disease mechanisms are very different (Verkerk et al., 1991, Huntington's Disease Collaborative Research Group, 1993). In fragile X, it is the very different risk of repeat expansion on maternal vs. paternal transmission that determines the anomalous inheritance pattern of the disorder (Fu et al., 1991). In Huntington disease, there is again a greater risk of allele expansion on transmission through a parent of one sex (male in this case). However, this manifests in a quantitative difference in the phenotype (particularly age of onset) rather than an altered inheritance pattern (Kremer et al., 1995, Squitieri et al., 2000). Neither of these examples of unstable mutation is considered to fall within the usual definition of genomic imprinting. Implicit in the concept of imprinting is that the parent-specific effects operate through postzygotic differences in gene expression or regulation that depend solely on parent of origin (Reik, 1989). In the case of the

trinucleotide disorders, in contrast, it is the genetic (rather than epigenetic) change that has occurred during gametogenesis that dictates the subsequent course of events.

1.1.7 CHARACTERISTICS OF IMPRINTED GENES

Several human chromosomes have had regions identified within them that are imprinted or are suspected to be so. This may be based on evidence from UPD cases, or from parent-of-origin-effects in single-gene disorders. It may also be deduced from the analysis of regions orthologous to those containing known imprinted genes in the mouse. More recently, some imprinted human genes have also been identified through systematic screens for differential expression or methylation (Hatada et al., 1991). To date, ~40 imprinted human genes are known, some of which are listed in Table 1.1.

Table 1.1. Mapped parent-of-origin effects in humans. Some of the genes identified as imprinted are listed. The chromosomal regions marked by an asterisk are further described in the text (modified from Morison and Reeve, 1998 and Morison et al., 2001).

CHROMOSOME	LOCATION	GENE/PARENT-OF ORIGIN-EFFECT
1*	1p36	p73 monoallelic expression has been identified in fetal thymus and pancreas; biallelic expression (or LOI) has been observed in renal and ovarian carcinomas (Mai et al., 1998, Chen et al., 2000a)
5	5q22-q31	*U2AFBPL* appears not to be imprinted in humans although its murine homologue (*U2afbp-re*) is imprinted
6*	6q25.3	*IGF2R* is imprinted in mice, and may be polymorphically imprinted in humans
	6q24-25	Transient neonatal diabetes, *ZAC1, HYMAI*
7*	7p11.2-p12	Silver-Russell syndrome, *GRB10, CIT1*
	7q32	*PEG1/MEST*
11*	11p15	Beckwith-Wiedemann syndrome; *H19, IGF2, KCNQ1, CDKN1C* and others – see below
	11q13 11q22.3-q23.3	Familial glomus tumours; tumour susceptibility is always inherited from carrier fathers
13	13q14	Serotonin receptor-2A (5-HTR2A) or *HTR2A*; expressed biallelically in normal tissues but in fibroblasts of some retinoblastoma patients is paternally expressed (Kato et al., 1996, Bunzel et al., 1998)
14	14q32	*PEG9/DLK1* is paternally expressed; *MEG3/GTL2* is maternally expressed (Kobayashi et al., 2000, Wylie et al. 2000)
15*	15q11-q13	Prader-Willi and Angelman syndromes. Among the imprinted genes identified are: *NDN, SNRPN, PAR4, IPW, UBE3A*
18	18q	The *IMPACT* gene appears not to be imprinted in humans though its murine homologue (*Impact*) is imprinted (Hagiwara et al., 1997, Okamura et al., 2000)
19*	19q13.4	*PEG3*
20*	20q13.3	*GNAS1*
X*	Xq13.2	*XIST*
	Xp11.23-Xqter	Turner syndrome patients with a maternally retained X (45, X^m) showed poorer verbal and higher order executive skills that those with 45,X^p

The genes described so far as imprinted may represent between 20 to 40% of their total number in the mammalian genome (Barlow, 1995). This estimate was based on the consideration that only 1 in 400 mouse loci analysed by phenotype and 2 in 263 mouse loci analysed by gene inactivation exhibit parental restriction. Other studies have shown similarly that only about 1 in 400 CpG islands in the mouse shows differential methylation (Shibata et al., 1995). Although these estimates may not in all cases reflect random sampling, extrapolation from a gene number of 60 000 suggests that between 100 and 200 mammalian genes may be imprinted.

Their growing number has allowed the delineation of some features that may be common to imprinted genes, as follows:

i) Differences in DNA methylation between the parental alleles
ii) Differences in histone acetylation between the parental alleles
iii) Differences in DNase sensitivity between the parental alleles
iv) Asynchronous replication of the two alleles
v) Imprinting regulation centres
vi) Repetitive sequences
vii) Antisense strand transcripts
viii) Clustering of imprinted genes

These aspects are considered further in the following paragraphs.

1.1.7.1 Differential methylation of imprinted genes

DNA methylation certainly has an important role in genomic imprinting. Not only do imprinted genes display different methylation patterns on their maternal and paternal alleles, correlated with differences in gene transcription, but these methylation states are also heritable through cell division, and are therefore capable of acting as an epigenetic mark (Shemer et al., 1996, Feil and Kelsey, 1997). The rather contrived neologism "methylome" has even been proposed recently, to refer to the complete set of DNA methylation modifications within a cell. This would include a description of

individual sites of DNA methylation and regional subchromosomal variations, not just global levels of methylation (Feinberg, 2001).

The DNA 5-cytosine methyltransferases are enzymes that catalyse the transfer of a methyl group from S-adenosyl methionine to the 5 position of cytosines, almost exclusively those within the dinucleotide sequence CpG (Bonfils et al., 2000). The 5-methyl groups do not affect base pairing, but can influence protein-DNA interactions by protruding into the major groove of the DNA helix, thus altering the binding affinity of sequence-specific transcription factors (Jones and Laird, 1999). To date, three distinct families of DNA methyltransferase genes have been identified, they have been termed *DMNT1/Dnmt1*, *DNMT2/Dnmt2* and *DNMT3/Dnmt3* (Xie et al., 1999).

The *DNMT1* gene maps to human chromosome 19p13.2-p13.3; it codes for a large enzyme, DNMT1, of 193 kDa (Bestor et al., 1988, Robertson et al., 1999). The mouse *Dnmt1* gene has 4 promoters and several spliced forms, including one, Dnmt1o, which is present only in the oocyte (Bestor, 2000, Bigey et al., 2000). DNMT1B is a minor isoform of the human protein, (resulting from a 48 nt insertion between exons 4 and 5 of the DNMT1 mRNA). It is ubiquitously expressed in human cells, but constitutes only 2-5% of the total amount of DNMT1 (Hsu et al., 1999, Bonfils et al., 2000).

The importance of Dnmt1 for CpG methylation of imprinted genes was directly demonstrated by the generation of mice carrying a targetted mutation in the *Dnmt1* gene. Embryos homozygous for the mutated allele exhibited genome-wide hypomethylation and loss of the allele-specific expression of the imprinted genes *Igf2* and *H19* (Li et al., 1992 and 1993, Reik et al., 1999, Sasaki et al., 2000). These mice failed to develop beyond the stage characteristic for normal day 9.5 embryos, and died at day 11 (Li et al., 1992). Further studies of *Dnmt1* were carried out by Jackson-Grusby et al. (2001). These authors conditionally inactivated *Dntm1*, demonstrating that the loss of *Dnmt1* from somatic cells causes p53-dependent apoptosis and also causes a global induction of gene expression, either directly or by

interaction with histone deacetylases and/or by changes in chromatin structure. Dnmt1 has a 5 to 30-fold preference for hemimethylated substrates, and it has therefore been attributed a role in maintenance of methylation patterns through successive rounds of semiconservative DNA replication (Reik et al., 1999, Bonfils et al., 2000).

From the point of view of imprinting, one of the most interesting characteristics of the mouse *Dnmt1* gene is that it has alternative sex-specific promoters and is subject to both transcriptional and posttranscriptional regulation within germ cells (particularly oocytes). Three alternative 5′ exons have been identified, which in 5′ to 3′ order are specific to oocytes (Dnmt1o, mRNA 5.1 kb), to somatic cells of both sexes (Dnmt1s, mRNA 5.2 kb), and to pachytene spermatocytes (Dnmt1p, mRNA 6.0 kb) (Mertineit et al., 1998). The most 5′ exon (the non-coding 157 nt exon 1o) directs synthesis of a mRNA that encodes an N-terminally truncated (but active) oocyte-specific Dnmt1o isoform (Gaudet et al., 1998, Mertineit et al., 1998).

Mertineit et al. (1998) demonstrated that *Dmnt1* (mRNA 5.2 kb) is present at high levels in spermatogonia and spermatocytes until the pachytene stage, when it falls to undetectable levels, but then an increase of the larger mRNA (Dnmt1p, 6.0 kb) is noted. In the oocyte, Dnmt1o accumulates during the growth phase, with a nuclear localisation, but later becomes cytoplasmic; it is found only at the oocyte cortex in ovulated oocytes, but re-enters the nucleus at the four-cell stage after fertilization (Gaudet et al., 1998, Mertineit et al., 1998).

The differential expression of these three DNMT1 isoforms may have important consequences in the context of genomic imprinting. From nuclear transplantation experiments, it is known that the nuclei of growth-stage oocytes are more effective in supporting the development of biparental embryos than the nuclei of less mature non-growing oocytes (Kono et al., 1996). This indicates that the biparental developmental potential of such embryos is acquired during oocyte growth. As Dnmt1o is present in both growing oocytes and preimplantation embryos, this isoform of the enzyme could have an important role in epigenetic modification of the

growing oocyte nucleus, and/or in the maintenance of imprinted methylation patterns during early development (Gaudet et al., 1998).

It became clear that DNA methyltransferases other than DNMT1 are involved in modifying mammalian DNA when further investigations of the *de novo* methylation and maintenance methylation mechanisms were carried out. As previously mentioned, the experiments of Li et al. (1992) showed that ES cells homozygous for the null *Dnmt1* mutation were viable. However, it later appeared that some residual Dnmt1 activity may have been present in these lines. Lei et al. (1996) therefore created a new truly null allele of *Dnmt1*, *Dnmt*^{c}. In homozygous mutant ES cells, they measured i) the levels of *de novo* methylation (by introducing a foreign DNA, the murine retrovirus MoMuLV) and ii) the maintenance of methylation. They observed stable low levels of both methylcytosine and methyltransferase activity. Furthermore, the methylation was maintained after more than 20 cell generations. They also found that the newly integrated MoMuLV DNA was almost completely methylated in both wild-type cells (controls) and mutated cells 8 days after infection. These results clearly indicated the existence of another enzyme(s) able both to maintain methylation of endogenous loci and to account for the *de novo* methylation activity observed in the ES cells.

A second *DNMT* gene, *DNMT2* was then identified by Yoder and Bestor (1998); its gene is on human chromosome 10p15 (Vilain et al., 1998, Yoder and Bestor, 1998). *Dnmt2* has 3 transcripts (1.6, 2.6 and 4 kb) the 1.6 kb transcript being the most abundant and ubiquitously expressed, although at very low levels. There is 81% amino acid sequence identity between human and mouse Dnmt2 enzymes (Okano et al., 1998). Interestingly, Dnmt2 is more similar to a putative DNA methyltransferase of the fission yeast (*S. pombe*) than to Dnmt1 (Yoder and Bestor, 1998), probably reflecting divergence very early in evolution.

By creating a *Dnmt2* null allele, Okano et al. (1998) demonstrated not only that ES cells homozygous for the mutation appeared to be normal in growth and morphology, but also that an endogenous virus maintained its methylation level and a newly

integrated retrovirus was *de novo* methylated. Their results suggested that Dnmt2 is not essential either to maintain methylation or to perform *de novo* methylation. In fact, the biological function of Dnmt2 remains undefined.

The *DNMT3* family consist of two related genes, *DNMT3A* (2p23) and *DNMT3B* (20q11.2). DNMT3A (912 aa) and DNMT3B (853 aa) share respectively 98% and 94% of amino acid sequence identity with their orthologous mouse enzymes. *DNMT3A* is ubiquitously expressed and has three major transcripts of 4.0, 4.4, and 9.5 kb; it is not known if all of these can encode functional proteins (Robertson et al., 1999, Xie et al., 1999, Ramsahoye et al., 2000). At least four different transcripts of *Dnmt3b* have also been identified (Robertson et al., 1999).

Dnmt3a can methylate not only CpG residues but also CpA and CpT. Given their close sequence similarity, it seems possible that Dnmt3b, too, may methylate non-CpG dinucleotides (Ramsahoye et al., 2000).

By infecting ES cells lacking either or both enzymes with a recombinant retrovirus ($MoMuLv^{sup}$-1), Okano et al. (1999) demonstrated that *Dnmt3a*$^{-/-}$or *Dnmt3b*$^{-/-}$ single mutants methylated proviral DNA normally, whereas the double mutant cells completely lacked *de novo* methylation activity. Furthermore, it was also demonstrated that both enzymes are required for *de novo* methylation during embryonic development. Doubly heterozygous mice (*Dnmt3a*$^{+/-}$or *Dnmt3b*$^{+/-}$) were normal and fertile, but *Dnmt3a*$^{-/-}$mice died early (4 weeks) after birth. The *Dnmt3b*$^{-/-}$ mutation was lethal *in utero*, the embryos presenting multiple malformations (such as neural tube defects) and growth impairment. Similar conclusions regarding the role of Dnmt3a as a *de novo* methyltransferase were reached by Lyko et al. (1999), using a transgenic *Drosophila* model.

Since Dnmt1 failed to *de novo* methylate retroviral DNA and various repetitive sequences in the Dnmt3a$^{-/-}$/Dnmt3b$^{-/-}$ ES cells and embryos, it would appear that by itself, Dnmt1 probably cannot initiate *de novo* methylation (Okano et al., 1999). It was also observed that the methylation of *H19* and *Igf2r* is retained in Dnmt3a$^{-/-}$

/Dnmt3b$^{-/-}$ ES cells, thus supporting the idea that the main role of Dnmt1 is maintaining parental methylation imprints once that they are introduced into the germline (Okano et al., 1999, Reik et al., 1999).

Mutations in *DNMT3B* cause the ICF syndrome (OMIM #242860), characterised by *I*mmunodeficiency, *C*entromere instability of chromosomes 1, 9 and/or 16, *F*acial anomalies (epicanthal folds, telecanthus, flat nasal bridge, macroglossia and mild micrognathia), and recurrent respiratory infections (Wijmenga et al., 1998). By homozygosity mapping Wijmenga et al. (1998) localized the ICF syndrome gene to a 9 cM region on chromosome 20q11-q13. The *DNMT3B* gene was also shown to map to 20q11.2 (Xie et al., 1999).

Okano et al. (1999) noted that one of the main characteristics of ICF syndrome (the hypomethylation of satellites II and III, major components of the constitutive heterochromatin) was similar to the hypomethylation of centromeric minor satellite repeats demonstrated in the *Dnmt3b*$^{-/-}$ mutant mice. This, added to the known genetic linkage data, suggested *DNMT3B* as a good ICF syndrome candidate gene. Homozygosity or compound heterozygosity for *DNMT3B* mutations were found in six ICF patients (Okano et al., 1999, Xu et al., 1999). It appears clear from these studies that DNMT3B has a specific role in the methylation of certain centromeric repeats, which in turn is important for the maintenance of chromosome stability.

1.1.7.2 Histone acetylation

Histone acetylation contributes to the formation of a transcriptionally competent environment by "opening" chromatin and allowing general transcription factors to gain access to the DNA. Conversely, histone deacetylation contributes to a "closed" chromatin state and transcriptional repression. Histone acetylation is carried out by histone acetyltransferases (HATs) and histone deacetylation by histone deacetylases (HDACs). Several HATs can also acetylate non-histone proteins such as p53 (Pazin and Kadonaga, 1997, Eden et al., 1998, Cheung et al., 2000).

The lysine residues at the N-terminal tails of the core histones are acetylated particularly within nucleosomes associated with transcribed genes. On the other hand, inactive genomic regions (such as the silent X chromosome and constitutive heterochromatin) are associated with hypoacetylated histones. It has been suggested that acetylation alters the nucleosome conformation either by decreasing the affinity of histones for DNA or by causing a change in the higher-order chromatin structure by interfering with internucleosomal contacts (Pedone et al., 1999).

The functional link between histone acetylation and DNA methylation has been found to include close interactions between the enzymatic machineries that regulate these two aspects of chromatin structure. Thus, the methylcytosine binding protein MeCP2 physically interacts with the transcriptional corepressor protein Sin3A, and in doing so recruits an HDAC to chromatin that contains methylated DNA (Amir et al., 1999). MeCP2 is a member of a family of proteins that mediate transcriptional regulation, and is mutated in patients suffering from Rett syndrome, an X-linked neurodegenerative disorder (Amir et al., 1999). MeCP2 has affinity for heavily methylated DNA rich in CpG. This direct link between methylation and histone acetylation explains previous observations that methylated DNA templates microinjected into cell nuclei are initially actively transcribed, but only until the DNA is packaged into chromatin (Jones et al., 1998, Tycko, 2000). The close functional relationship between methylation and HDAC activity is further shown by the fact that DNMT1 forms transcriptionally repressive multiprotein complexes with both HDAC1 and HDAC2 (Fuks et al., 2000, Robertson et al., 2000, Rountree et al., 2000). The interacting effects of histone deacetylation and DNA methylation are also shown by the need for a combination of both trichostatin A (a potent HDAC inhibitor) and 5 aza-dC (a methylation inhibitor) to reactivate the silent paternal *H19* allele in culture (Pedone et al., 1999, Cheung et al., 2000).

1.1.7.3 DNase sensitivity

A DNaseI hypersensitive site corresponds to a region where nucleosomes are either absent or are partially disrupted due to the binding of non-histone proteins (Feil and

Khosla, 1999). In many genes, DNaseI hypersensitivity marks the sites of regulatory sequences such as enhancers or locus control regions, reflecting the altered chromatin structure that results from binding of transcriptional regulators to such sites (Schweizer et al., 1999).

Not surprisingly, altered DNaseI sensitivity accompanies the differential expression state of the alleles of some imprinted genes, too. In the case of *U2af1-rs1*, the silent maternal allele is hypermethylated in a CpG-rich region in which the expressed paternal allele is unmethylated. DNase I hypersensitivity of a putative promoter region, indicative of an "open" chromatin conformation, was seen only on the paternal allele of this gene (Shibata et al., 1996).

DNase I hypersensitivity has also been assessed within the differentially methylated regions (DMRs) of the mouse *H19* locus (Hark and Tilghman, 1998). Two regions of nuclease hypersensitivity were identified on the maternal chromosome, which corresponded to the most heavily methylated regions on the paternal allele. The more distal hypersensitive region (or HS1) is <500 bp in length and is centred at –3.8 kb from the *H19* promoter, whereas the promoter-proximal region (HS2) extends from approximately –2.0 to -2.7 kb (Hark and Tilghman, 1998, Ben-Porath and Cedar, 2000). Regulation of *H19* is further discussed in section 1.1.10.3.

1.1.7.4 Asynchronous replication, homologous pairing and meiotic recombination frequencies of imprinted genes

1.1.7.4.1 Replication timing

Replication timing can be assessed by *in situ* hybridisation to interphase nuclei. Nuclei in which a given gene has not yet replicated show two single hybridisation dots, whereas those in which both alleles have already replicated reveal two sets of doublets. Thus in an asynchronous cell-cycle population, the relative time of replication can be measured, since the higher the proportion of double-spotted nuclei, the earlier that gene replicates in S-phase. Most genes replicate synchronously on the two chromosomal homologues, with only ~10% of nuclei showing an asynchronous

(single-double) hybridisation pattern (Kitsberg et al., 1993). In contrast, 25 to 35% of nuclei display asynchronous replication of the imprinted genes *Igf2*, *Igf2r* and *H19*, in each case the paternal allele being early replicating (Kitsberg et al., 1993, Greally et al., 1998). It has been suggested that replication timing may be controlled by cell-cycle dependent trans-acting factors that bind to recognition sites on the DNA exclusively at the time they undergo replication. Once an initial allele-specific target site for such factors has been set up in the gametes, the differential replication timing will be preserved, representing a primary epigenetic difference between the parental alleles (Simon et al., 1999).

1.1.7.4.2 Homologue pairing

Homologue pairing refers to the preferential spatial association of homologous loci in meiotic and somatic cells. In studies of imprinted genes, it has been demonstrated that an association between the two homologous 15q11-q13 chromosomal regions occurs during the late S-phase of the somatic cell cycle. Cells from PWS and AS patients were demonstrated to be deficient in this association, suggesting that normal imprinting may also involve mutual recognition and preferential association between maternal and paternal chromosomes (LaSalle and Lalande, 1996). The precise nature of the interactions that occur between the homologous chromosomal regions, however, remain unclear.

1.1.7.4.3 Recombination frequencies

Large sex-specific differences in meiotic recombination frequencies have been observed in imprinted regions. The 11p15.5 and 15q11-q13 regions both recombine with very high frequency during male meiosis and very low frequency in female meiosis. It has been suggested that this phenomenon may be related to region-specific differences in accessibility of DNA to the recombination machinery (Paldi et al., 1995).

1.1.7.5 Repetitive sequences in imprinted genes

Neumann et al. (1995) described the presence of direct tandem repeats embedded in the CpG-rich differentially-methylated regions of imprinted genes. Although these repeats tend to be G-rich, no specific recurring sequence motifs have been identified between different genes. The repeat units vary in length and location with respect to their genes. The exact role of these sequences is unknown, but it has been suggested that they may form specific secondary structures that attract *de novo* methylation either in the gametes or soon after fertilisation.

1.1.7.6 Antisense transcription at imprinted loci

Antisense transcripts have been described within or closely associated with several imprinted genes. It is believed from various lines of evidence that these antisense transcripts may play a role in regulating the imprinted genes with which they are associated. In general, these antisense RNAs do not have an open reading frame and hence probably exert their effects directly within the nucleus at the RNA level.

The antisense transcript that is best characterised functionally is the one associated with the mouse *Igf2r* gene, named *Air*. This originates from a differentially methylated promoter region in intron 2 of *Ig2r*, and is transcribed exclusively from its unmethylated paternal allele. Since *Igf2r* expression is maternally monoallelic, it was suspected that *Air* is involved in the suppression in *cis* of transcription from the sense *Igf2r* promoter. This idea received direct support from transgenic experiments in which deletion of the *Air* promoter region result in biallelic *Igf2r* expression (Wutz et al., 1997). Thus, transcription of *Air* across the paternal *Igf2r* promoter may have the effect of somehow "occluding" the sense gene promoter, perhaps locally inactivating the chromosome by physically "coating" the chromatin (as has been proposed for the X chromosome inactivation, see below). Alternatively, the *cis* suppressive effect may result from sharing of transcription factors that are in limiting supply or from competition in *cis* for shared enhancers (Reik and Constancia, 1997, Constancia et al., 1998).

Several imprinted antisense transcripts (*Igfas a, b, c*) have also been described in the upstream region of the *Igf2* gene. However, in contrast to the *Igf2r* situation, these are expressed, at low levels, from the same allele as *Igf2*. Their role is unclear, although it has been suggested that the overlapping antisense transcripts may influence the activity of sense promoters, and/or may be involved in regulating *Igf2* in a tissue-specific manner (Moore et al., 1997).

1.1.7.7 Clusters of imprinted genes and regulation by imprinting centres

One feature that has repeatedly emerged from studies of the chromosomal regions containing imprinted genes is that such genes tend to be grouped together in clusters (Lee et al., 1999b). Detailed analysis of these has revealed complex regulatory mechanisms that often indicate coordinated regional control of the imprinted domain. The two most intensively studied imprinted gene clusters are on human chromosomes 11 and 15 (with syntenic homology to mouse chromosome 7, distal and central respectively) (Nicholson et al., 2000). However, other regions of clustering of imprinted genes have been identified on human chromosomes 7 and 14, and there are also some other examples in the mouse (Caspary et al., 1998).

A related and important feature of both of the above two imprinted chromosomal regions is the so-called imprinting centre or IC (these are discussed further in sections 1.1.9 and 1.1.10). The IC can be conceptualised as a *cis*-acting control element that exists in two alternative structural forms. Imprint switching is postulated to occur during gametogenesis when the maternal IC allele is set up in one conformation, whereas the paternal allele adopts the alternative structure. Once established, each of these structures is maintained through fertilisation and embryonic development. Following implantation, the IC can act as a regional organiser which sets up imprinted patterns of methylation, chromatin structure and gene expression across large genomic intervals (Ben-Porath and Cedar, 2000).

Our understanding of imprinting centres depends largely on studies of clinical disorders in which imprinting of one of the above-mentioned clusters (15q and 11p)

has been disturbed. Therefore, before discussing models that have been proposed to explain the mechanism of operation of the IC, it is necessary to review in some detail each of these two clusters, including the individual genes they contain and relevant aspects of the pathology associated with them. This is done in the following sections.

1.1.8 THE CLUSTER OF IMPRINTED GENES IN 15q11-q13

Much recent effort has gone into the identification and analysis of transcripts within the large imprinted region of chromosome 15q11-q13 (see for example Lee and Wevrick, 2000, Meguro et al., 2001). This predominantly paternally expressed domain is bounded on the centromeric side by the marker *D15S1035*; all transcripts centromeric to this are apparently biallelically expressed. On the telomeric side, the paternally expressed domain is bounded by the gene *UBE3A*, with all transcripts distal to this gene again being expressed from both alleles (Figure 1.1). *UBE3A* itself, as discussed below, is maternally expressed in a tissue-specific pattern. As already described, Prader-Willi syndrome (PWS) and Angelman syndrome (AS) are both associated with anomalies of the imprinting of this region.

Figure 1.1. Gene map of human chromosome 15q11-q13 (not to scale). The map indicates the direction of the centromere (cen) and telomere (tel), and identifies some of the genes within this region for reference. These are shown as dark circles (paternally expressed), grey circles (maternally expressed) or open squares (nonimprinted). An imprinting control centre (IC) associated with the 5′ end of the bicistronic *SNRPN-SNURF* locus is indicated. The large arrow below this indicate the regional imprinting control exerted from the IC. The small arrows show overlapping sense and antisense transcription. The open block arrow indicates the extent of the region frequently deleted in PWS and AS. The locations of the markers *D15S1035*, *D15S174* and *D15S165*, referred to in the text, are indicated. The dark arrow indicates the critical region implicated in PWS and the grey block arrow indicates the region implicated in AS (Modified from Ji et al., 1999, Lee and Wevrick, 2000, De los Santos et al., 2000, Nicholls, 2000).

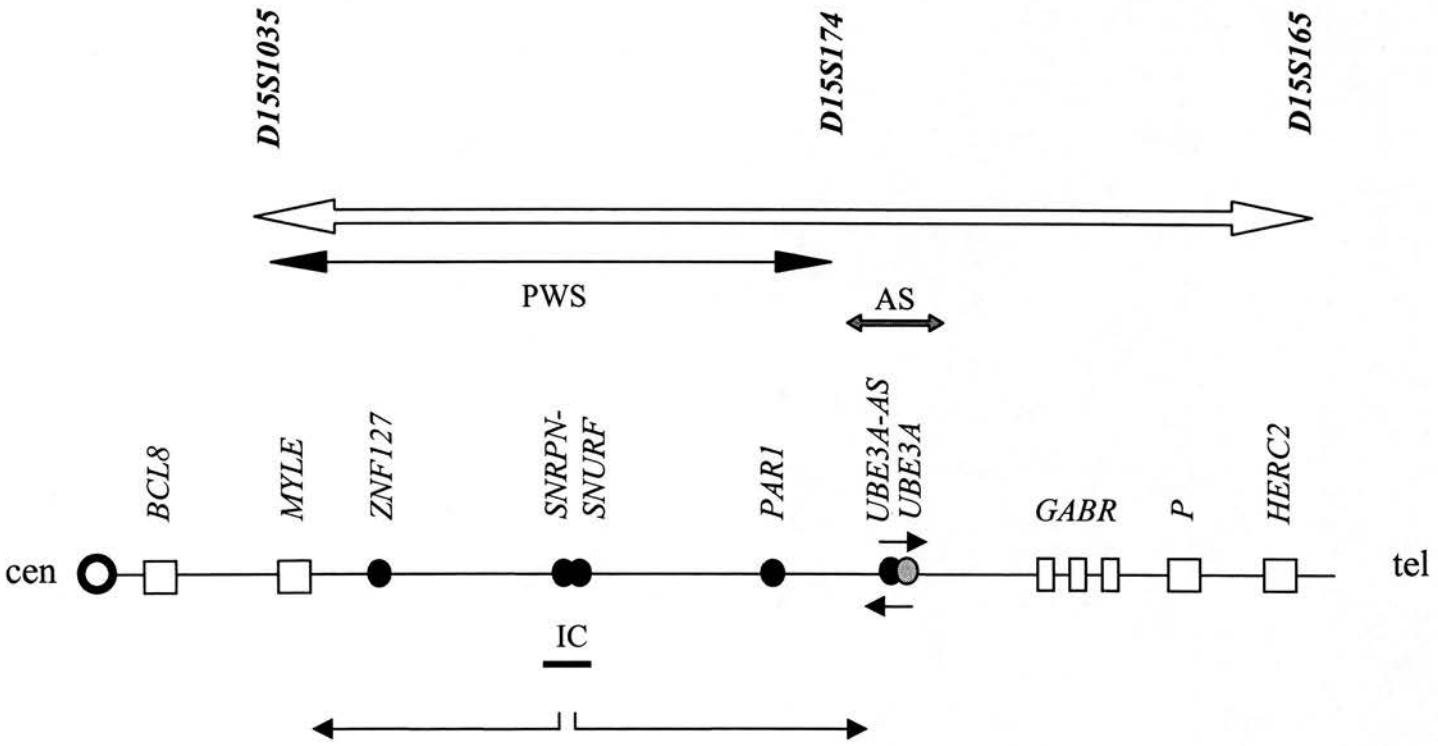


Figure 1.1. Gene map of human chromosome 15q11-q13 (not to scale).

1.1.8.1 The Prader-Willi and Angelman syndromes

PWS (OMIM #176270) and AS (OMIM#105830) are distinct clinical entities that were first described in 1956 and 1965 respectively (Prader et al., 1956, Angelman, 1965, Bower and Jeavons, 1967, Cassidy, 1997). It was only decades later that cytogenetic and then also molecular evidence converged to link the two conditions. PWS and AS are considered the classical manifestations of genomic imprinting in man since they result from the absence of either maternal (AS) or paternal (PWS) genomic contributions from 15q11-q13. This, in turn, may result from various different mechanisms affecting one or more genes in this region (Ledbetter et al., 1981, Smith et al., 1992, Chan et al., 1993, Sutcliffe et al., 1994, Smith et al., 1997, Matsuura et al., 1997).

1.1.8.2 Prader-Willi syndrome

PWS has been intensively studied over a period of many years; it was the first example of a microdeletion syndrome identified by high resolution chromosome analysis, the first recognised human disorder of genomic imprinting, and also the first disorder recognised as resulting from UPD (Butler et al., 1986, Cassidy, 1997).

PWS has an incidence of 1 in 10,000-15,000, occurring in both sexes and all races (Cassidy, 1997, Christian et al., 1998). The clinical characteristics include infantile hypotonia, hypogonadism, and later in childhood hyperphagia and obesity. Patients have a narrow bifrontal diameter, almond shaped palpebral fissures, narrow nasal bridge, small narrow hands, short stature, developmental delay, and behavioural problems. A third of patients also present hypopigmentation (Cassidy, 1997, Couper and Couper, 2000).

Between 60 and 70% of PWS cases are due to a ~4 Mb deletion of 15q11-q13 on the paternally-derived chromosome. 28% of cases are due to maternal UPD. These two common mechanisms are functionally equivalent in resulting in lack of a paternal 15q11-q13 contribution. A small but biologically very interesting group of patients

(less than 2%) have structurally normal, biparentally-derived chromosome 15s, but have an abnormality of the imprinting process which results in non expression of the paternal genes within the PWS critical region (Cassidy, 1997).

1.1.8.3 Angelman syndrome

AS has an estimated incidence of 1 in 20,000. It is characterised by severe mental retardation, absent speech, abnormal laughter and ataxic gait, seizures or abnormal EEG, microcephaly, brachycephaly, and a characteristic facies with macrostomia and prognathism (Williams et al., 1988, Meijers-Heijboer et al., 1992, Smith et al., 1997, Fridman et al., 2000). Bower and Jeavons (1967) likened the easily provoked and prolonged paroxysms of laughter and ataxic jerky movements to the movements of puppets; hence this condition was for a while referred to as the "happy puppet" syndrome (Magenis et al., 1988).

Like PWS, AS patients may be pathogenetically classified into one of the following groups (Table 1.2) (Cassidy, 1997):

TYPE	MECHANISM	PROPORTION OF CASES	METHYLATION	RECURRENCE RISK
Ia	~4-Mb interstitial maternal del(15q11-q13)	65%-75%	Abnormal	Extremely low
Ib	Unbalanced translocation or inherited interstitial deletion	<1%	Normal or abnormal	Significant
IIa	UPD (maternal deficiency) with normal parental chromosome	3%-5%	Abnormal	Extremely low
IIb	UPD with predisposing parental translocation	<1%	Abnormal	Significant
IIIa	Imprinting mutation with deletion of IC	3%-5%	Abnormal	Significant
IIIb	Imprinting mutation without detectable deletion of IC	3%-5%	Abnormal	Low
IV	Point mutation in *UBE3A*	4%-6%	Normal	Significant
V	AS phenotype with no identifiable molecular abnormality	10%-14%	Normal	Occurs rarely

Table 1.2. Classification of AS according to underlying cause. The term "abnormal" is used to indicate that only the paternal, unmethylated pattern is seen on analysis of the *SNRPN* promoter region (as in Jiang et al., 1999).

AS may result from deletions, UPD or imprinting mutations, but in addition in some cases point mutations in the *UBE3A* gene (including truncating and missense mutations) have been described as the only abnormality. Unlike PWS therefore (which is generally considered a contiguous gene syndrome), AS appears to be a single-gene disorder. There remains, however, a substantial number of patients with an AS phenotype (>10%) in whom none of these molecular or cytogenetic alterations can be identified. Some of these may be attributable to misdiagnosis or a phenocopy (Matsuura et al., 1997, Fang et al., 1999).

Although, as noted in Table 1.2, the recurrence risk for cytogenetic deletion cases of AS is very low, at least one case of maternal germ-line mosaicism for the AS-associated 15q deletion has been reported. In this AS family with recurrence of del(15), the mother's chromosomes 15 were structurally normal, but the two patients and their unaffected brother all shared an identical maternally-derived haplotype distal to the deletion region, suggesting that they had indeed inherited the same grandpaternal chromosome 15 from the mother (Kokkonen and Leisti, 2000).

1.1.8.4 The 15q11-q13 chromosomal region

A 15q11-q13 deletion was first described in PWS patients in 1980 (Ledbetter et al., 1980). This was also confirmed by Butler et al. (1986), who by analysing chromosome 15 short arm and satellite region variants demonstrated that the del(15q) was paternal in origin and also that the chromosomes of both parents were structurally normal. Later studies by Nicholls et al. (1989) of two patients who did not have cytogenetic deletions, demonstrated that in these cases the PWS was due to a maternal uniparental heterodisomy, the first time that this situation had been associated with a human genetic disease.

The cytogenetic and molecular histories of PWS and AS intersected when several patients whose phenotype was compatible with AS were reported to have a deletion also involving 15q11-q13 (Magenis et al., 1987 and 1988, Williams et al., 1988). Molecular analysis indicated that in contrast to PWS patients, in these AS cases it was always maternal markers that were absent (Knoll et al. 1989, Smith et al., 1992). Magenis et al. (1990) further showed that the extents of the deleted regions in PWS and AS were essentially the same, implying that differential expression of one or more genes from the maternal and paternal homologues must underlie the different phenotypes of the two conditions. A similar conclusion was reached from observations on a rare example of a familial deletion, described by Greenstein (1990). One individual in this family had PWS, while a cousin had AS. The father of the child with PWS and the mother of the child with AS were brother and sister,

consistent with the known parental origins of sporadic cases and strongly implying that the deletions were identical in extent in both patients.

1.1.8.5 Genes identified in the 15q11-q13 region

Several groups have exerted efforts towards gene identification in the PWS/AS deletion region of 15q11-q13, Genes and transcripts have been scrutinized to differing extents to establish their imprinted or non-imprinted patterns of expression. So far ~15 genes and more than 20 anonymous ESTs have been mapped to this region. In this respect, it is worth considering that according to a notional genome-wide average gene density, a region of this size might be expected to contain ~100 genes (Figure 1.1) (Lee and Wevrick, 2000).

1.1.8.6 Genes within the PWS minimal region

Any gene considered to be a candidate for PWS should lie between the markers *D15S1035* and *D15S174* (Figure 1.2). However, the PWS critical region is very large, and single-gene mutations or kb-size deletions have not been identified (with the exception of the IC mutations, see below). It therefore seems likely that PWS is a contiguous gene syndrome (Mann and Bartolomei, 1999, Nicholls, 1999). The examples below of individual genes that map to this region are discussed in a centromeric to telomeric order:

Figure 1.2. Partial gene map of human chromosome 15q11-q13 (not to scale). A) The directions of the centromere (cen) and telomere (tel) are indicated. The block arrow identifies the markers referred to in the text that enclose the PWS region. Some of the genes that map within this region are indicated for reference. B) The genes that map to this region (and that are further discussed in the text) are indicated as before (Figure 1.1): dark circles (paternally expressed), grey circles (maternally expressed), open squares (nonimprinted). EST *SGC32610* is represented as a grey square as its expression was found to be biallelic in fibroblast cell lines and brain but maternally biased in lymphoblastoid cell lines. (Modified from Ji et al., 1999, Lee and Wevrick, 2000, De los Santos et al., 2000, Nicholls, 2000.)

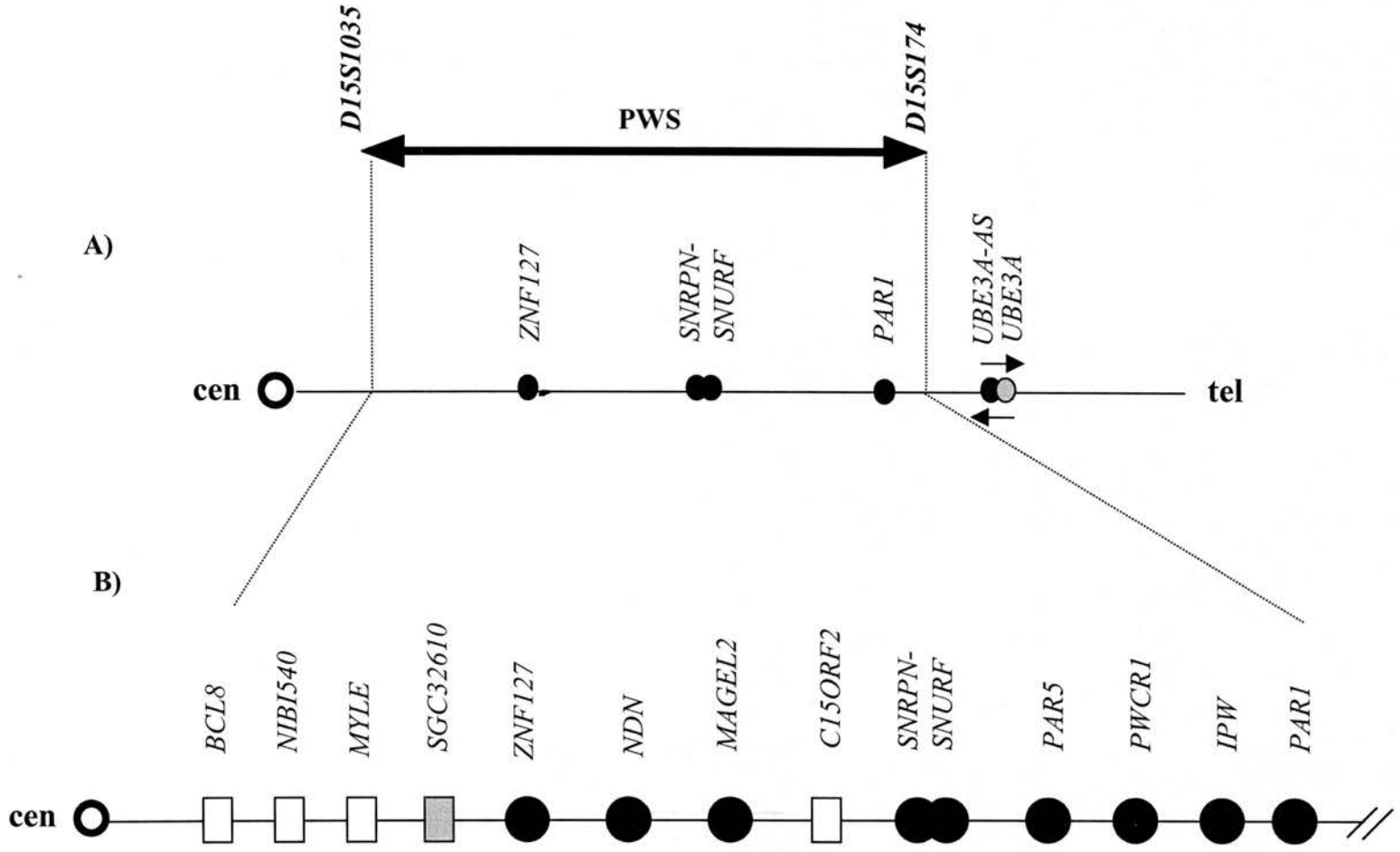


Figure 1.2. Partial gene map of PWS region of human chromosome 15q11-q13 (not to scale)

i) The *BCL8* gene is centromeric to the border of the critical region (it lies beyond the *D15S1035* marker limit; Figure 1.1). The *BCL8* region is a recurrent site of rearrangements in diffuse large-cell lymphomas (DLCL). This gene has not yet been completely characterised but two transcripts have been identified, a more abundantly expressed of 2.6 kb and a 4.5 kb transcript (Dyomin et al., 1997). RT-PCR analyses have demonstrated that *BCL8* is normally expressed in human testis and prostate and frequently expressed in DLCL but interestingly, this gene is not normally expressed in lymphoid tissues. *BCL8* is thus the only gene known (so far) to be ectopically activated in lymphomas, and is presumed to have a role in the pathogenesis of this disease (Dyomin et al., 1997). The imprinted status of *BCL8* itself has not been investigated. However, at least three ESTs (*NIB1540*, *SHGC13938* and *A00B45*) that map between *BCL8* and *D15S1035* and also one EST (*SHGC15126*) corresponding to the poorly-characterised *MYLE* gene are all biallelically expressed in lymphoblast cell lines, fibroblast cell lines and brain (see Figure 1.2). It therefore seems likely that *BCL8* is also non-imprinted (Lee and Wevrick, 2000).

ii) The EST *SGC32610* is telomeric to *MYLE* and maps within the PWS region. It is biallelically expressed in fibroblast cell lines and brain but is preferentially maternally expressed in lymphoblastoid cell lines (Lee and Wevrick, 2000). This is somewhat reminiscent of *UBE3A* (at the opposite end of the imprinted domain), for which incomplete imprinting in brain has been shown (see below).

iii) Further telomeric to *BCL8* is the intronless gene *ZNF127* (Figures 1.1 and 1.2). This encodes a protein with a RING C_3HC_4 zinc-finger (C, Cys; H, His) and multiple C_3H zinc-finger motifs. The RING zinc-finger domain was first identified in the *RING1* gene within the MHC region of chromosome 6p21.3 (Lovering et al., 1993). The C_3HC_4 RING-finger motif (Cys-X_2-Cys-$X_{(9-39)}$-Cys-$X_{(1-3)}$-His-$X_{(2-3)}$-Cys-X_2-Cys-$X_{(4-48)}$-Cys-X_2-Cys, where X is any amino acid) has been identified in more than a hundred different proteins with a diverse range of functions. The RING finger is likely to function in protein-protein interactions and many RING proteins may mediate formation of macromolecular complexes (Saurin et al., 1996, Jong et al., 1999a and b).

The C_3H motif has the consensus Cys-$X_{(1-2)}$-Tyr/Phe-$X_{(2-5)}$-Gly-$X_{(1-2)}$-Cys-X2-Gly-X_2-Cys-X-Tyr/Phe-X-His. (Jong et al., 1999 a and b). Although the exact function of the C_3H domain has not been elucidated, Carballo et al. (1998) showed that tristetraprolin, the prototype C_3H zinc finger protein, inhibited the production of TNF-α by destabilizing its messenger RNA. This suggested that C_3H proteins as a class may be RNA-binding proteins.

Human and mouse ZNF127 constitute a separate subfamily of RING-finger proteins with both C-terminal RING-finger and C_3H motifs. As ZNF127 has three C_3H motifs, it may mediate RNA binding as part of a ribonucleoprotein complex in conjunction with the RING-finger (Jong et al., 1999 a and b).

Differential methylation of *ZNF127* has been described; in brain its CpG island is completely unmethylated on the expressed paternal allele, but methylated on the maternal allele (Driscoll et al., 1992; the cDNA DN34, referred to in this paper, was later shown to correspond to *ZNF127*, Jong et al., 1999b). The 3′ end of *ZNF127* is located within an intron of *ZNF127AS*, an antisense transcript that is also paternally expressed (Jong et al., 1999a and b).

iv) The intronless *NDN* gene is paternally expressed, and encodes the DNA-binding protein necdin. *NDN* is expressed in multiple tissues and in virtually all postmitotic neurons from early neurogenesis until adult stages (Nakada et al., 1998, Jay et al., 1997, Sutcliffe et al., 1997, Gerard et al., 1999). Necdin appears to act as a neuronal growth suppressor that is functionally similar to the retinoblastoma tumour suppressor protein (Rb) (although it is structurally dissimilar). Like Rb, necdin can suppress cell proliferation by interacting with the transcription factor E2F1, which promotes cell-cycle progression. Necdin has also been demonstrated to suppress the growth of Rb-deficient osteosarcoma cells (SAOS-2). Thus, necdin might complement Rb function in order to prevent postmitotic neurons from resuming cell division (Taniura et al., 1998, Nakada et al., 1998).

v) The *MAGEL2* gene (or *NDNL1)* lies 41 kb distal to *NDN.* It encodes a 529 amino acid protein with 51% sequence identity to necdin. MAGEL2 contains an additional 188 amino acids at the N-terminus compared with NDN and does not have a proline-rich, acidic N-terminus, as necdin does. It is expressed only from the paternal allele in brain. The murine orthologue *Magel2* is also imprinted, with paternal-only expression predominantly in late developmental stages; the RNA expression pattern of *Magel2* in the developing nervous system overlaps with that of *Necdin* (Lee et al., 2000).

vi) The intronless gene *C15ORF2* (*c*hromosome 15 *o*pen *r*eading *f*rame *2*) spans 7.5 kb. Its 3.5 kb open reading frame encodes a predicted 1156 amino acid protein. Despite being flanked by two imprinted genes (*MAGEL2* and *SNRPN*), it is biallelically expressed in adult testis (the only tissue where it is expressed) and therefore very probably does not contribute to the PWS phenotype (Farber et al., 2000).

vii) The *SNRPN* gene is paternally expressed and encodes a small (240 aa) nuclear spliceosomal ribonucleoprotein. There is a processed pseudogene, *SNRPNP1*, on human chromosome 6 (Leff et al., 1992, Ozcelik et al., 1992, Gray et al., 1999). There is some confusion in the literature over exon numbering within *SNRPN.* Eight exons were originally described; when 2 new exons were later discovered they were initially referred to as –1 or α and 0 or β, but have since been renumbered as exons 1 and 2. Also, several alternative 5′ exons of *SNRPN* have been identified. These are part of a transcript known as *BD*, and are of particular relevance because they lie within the imprinting centre (IC) that regulates this region (Dittrich et al., 1996). The most recent data indicates that the genomic structure of *SNRPN* is even more complex; the identification of new exons at its 3′ end brings the total number to date to 20. These eight new 3′ end exons do not appear to have any coding potential (Wirth et al., 2001).

Using phylogenetic analysis of *SNRPN* mRNA in different eutherian mammals, Gray et al. (1999) identified a second highly conserved coding sequence within this gene,

which they termed *SNURF* (*SNRPN* *u*pstream *r*eading *f*rame), that encodes a 71 amino acid protein. Thus it appears that *SNRPN* is a bicistronic gene.

By using RT-PCR and parent-specific DNA methylation analysis, Glenn et al. (1996) demonstrated that *SNRPN* is imprinted. It is differentially methylated at the CpG island associated with its exon 1 (originally $-1/\alpha$), being methylated on the repressed maternal allele; another CpG island identified in intron 7 was demonstrated to be preferentially methylated on the expressed paternal allele (Kubota et al., 1997, Cassidy, 1997). Methylation analysis of the *SNRPN* exon 1 (α) was also performed by Beuten et al. (1996) and is now commonly used for the molecular diagnosis of PWS.

Recent analyses of chromosomal translocations in PWS patients have identified a new breakpoint translocation region that maps 70-80 kb distal to the *SNURF-SNRPN* gene. These translocations disrupt some of the new 3′ exons, but it has also been hypothesised that they may disrupt the *HBII-85* sequence (see following paragraphs) or even affect an as yet unknown gene in this region (Wirth et al., 2001).

Meguro et al. (2001) carried out an extensive study of the allelic expression of 118 cDNA clones identified in the GeneMap database as lying within the PWS/AS region delimited by the genetic markers *D15S1035* and *D15S165*. They demonstrated the presence of a number of transcripts that while lacking protein coding potential, were expressed only from the paternal allele. Furthermore, within this region a large direct repeat (DR) cluster was also identified. This contains box C, D′ and D sequence motifs characteristic of a particular class of snoRNAs (small nucleolar RNAs). There are over 150 known snoRNAs, many of which are believed to have specific roles in the posttranscriptional modification of rRNA prior to its incorporation into the assembling ribosome. The C/D box and H/ACA box snoRNAs are two families characterized by specific sequence motifs, thought to have roles in ribose 2′-O-methylation and in pseudouridylation, respectively (Lowe and Eddy, 1999, De los Santos et al., 2000).

Three brain-specific C/D box snoRNAs and one H/ACA-box snoRNA were also identified by Cavaille et al. (2000), in both human and mice. The three C/D box snoRNA genes were mapped to the PWS region of human chromosome 15 and were given the names *HBII-13, HBII-52, HBII-85* (and *MBII-13, MBII-52, MBII-85* for the mouse orthologues). These snoRNAs were again paternally expressed, as judged from their absence from brain tissue of a PWS patient. It has therefore been considered that the DR cluster that encodes these various snoRNA transcripts may contribute to the phenotype of PWS (Meguro et al., 2001).

viii) *PAR1* and *PAR5* (for Prader-Willi/Angelman region) are two paternally expressed non-coding transcripts identified in fetal brain cDNA libraries (Sutcliffe et al. 1994). *PAR1* and *PAR5* are located within 300 kb telomeric to the *SNRPN* gene. Between these two genes (among other ESTs) map the *PWCR1* and *IPW* genes (see below), which are also non-coding (De los Santos et al., 2000). Small deletions of a region that includes *SNRPN* exon 1 (exon α) were found to result in loss of expression not only of *SNRPN*, but also of *PAR1* and *PAR5*, thus defining the presence of a paternal imprinting control region (see below) (Sutcliffe et al., 1994).

ix) *PWCR1* (for Prader-Willi chromosome region 1), is an intronless gene without an open reading frame. The most abundant of its several alternative transcripts is a short (140 nt) RNA which is well conserved between human and mice and represents a member of the C/D box snoRNA family (see above). Like these other transcripts, *PWCR1* is not expressed in PWS patients with either the classical deletion or an IC microdeletion, indicating that it requires the action of the IC regulatory element for its expression (De los Santos et al., 2000).

x) The next gene in a telomeric direction is *IPW* (maps ~250 kb distal to *SNRPN*). This is widely expressed, only from the paternal allele, but again does not appear to encode a protein (Glenn et al., 1996, Wevrick and Francke, 1997).

1.1.8.7 Genes within the AS critical region

The AS region lies between the markers *D15S174* and *D15S165* (Figure 1.3). The most centromeric gene identified within this interval is *UBE3A*.

Figure 1.3. Map of part of human chromosome 15q11-q13 (not to scale) showing the AS region. The map indicates the direction of the centromere (cen) and telomere (tel) and identifies some of the genes within this region (discussed in the text). The mono- or bi-allelic pattern of expression of the genes is indicated as before (Figures 1.1 and 1.2). The $GABA_A$ receptor genes are indicated as dark squares, as conflicting evidence has identified them as biallelically and/or paternally expressed (see text). The positions of duplicated sequences related to *HERC2* in this region are indicated by grey ovals. The open block arrow indicates the extent of the region frequently deleted in AS. The locations of the markers *D15S174* and *D15S165*, referred to in the text, are indicated. The grey block arrow indicates the AS critical region (see text). Modified from Ji et al., 1999, Lee and Wevrick, 2000, De los Santos et al., 2000, Nicholls, 2000.

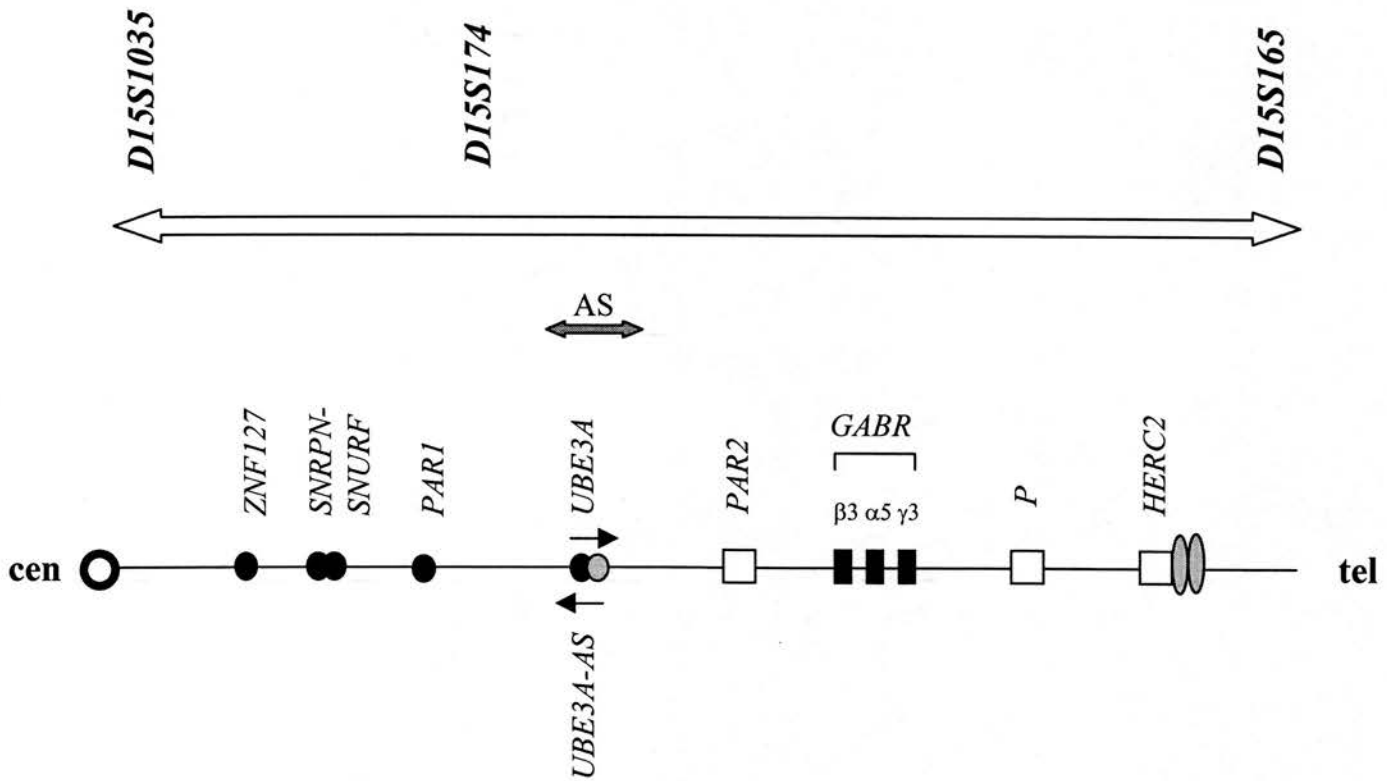


Figure 1.3. Partial gene map of human chromosome 15q11-q13 showing the AS region (not to scale)

i) *UBE3A* encodes the E6-AP ubiquitin-protein ligase (see below). Its 16 exons (six of which make up the 5′-UTR) span ~120 kb. Two processed pseudogenes, *UBE3AP1* and *UBE3AP2*, with 90-95% nucleotide sequence identity to *UBE3A*, have been mapped to chromosomes 2 and 21 respectively. At least five alternatively spliced mRNA forms of *UBE3A* have been identified (Kishino et al., 1997, Kishino and Wagstaff, 1998).

E6-AP is a member of the ubiquitin-protein ligase family that is defined by the presence of a carboxy terminal "hect" (homologous to the E6-AP carboxyl terminus) domain (Nakada et al., 1998). The ubiquitination process involves four different classes of proteins that act together to target selected proteins for degradation (Jiang et al., 1999).

UBE3A is biallelically expressed in fibroblasts and lymphoblasts. However, this gene is imprinted in certain regions of the brain in both humans and mice, showing maternal monoallelic expression (Albrecht et al., 1997, Rougeulle et al., 1997, Vu and Hoffman, 1997). Chromosomal inversions, translocations and mutations affecting *UBE3A* have all been described in AS patients that have a normal methylation pattern at the *SNRPN* locus (Michaelis et al., 1995, Greger et al., 1997, Kishino et al., 1997, Kishino and Wagstaff, 1998, Matsuura et al., 1997, Fang et al., 1999).

Besides *UBE3A*, there are other transcripts that have been identified in close proximity. These are an intronless transcript (transcribed in the same direction as *UBE3A*) of 3.5 kb in length, and an antisense transcript with its 5′ end 6.5 kb upstream of the *UBE3A* stop codon, crossing the 3′ end of *UBE3A*. The intronless sense transcript has the same imprinting pattern as *UBE3A*, while in contrast the antisense transcript is oppositely imprinted, with preferential expression from the paternal allele in brain (Rougeulle et al., 1997).

As for *Igf2r/Air*, the presence of the antisense transcript poses questions regarding its role in *UBE3A* regulation. For example, a transcript competition model proposes that

in tissues such as brain, where the antisense transcript is monoallelically (paternally) expressed, it prevents expression in *cis* of *UBE3A*, which can therefore only be expressed from the antisense maternal chromosome. In other tissues, the antisense transcript is not expressed and *UBE3A* can be transcribed from both alleles (Rougeulle et al., 1998). Unlike the *Igf2r/Air* locus, however, this model has not been experimentally tested at this locus.

ii) The *PAR2* gene, for Prader-Willi/Angelman region gene 2 (marker *D15S225E*), is telomeric to *UBE3A* and has been described as not imprinted (Nakao et al., 1994).

iii) The genes encoding the γ-aminobutyric acid (GABA) type A receptor subunits, β3 (*GABRB3*), α5 (*GABRA5*) and γ3 (*GABRG3*) span 800 kb of the distal part of the PWS-AS region. GABA, the principal inhibitory neurotransmitter in the central nervous system, exerts its effects through these $GABA_A$ receptors. Conflicting studies have described these genes as either imprinted (Meguro et al., 1997) or not imprinted (Gabriel et al., 1998). In the mouse, Nicholls et al. (1993) had concluded that *Gabrb3* was non-imprinted. Meguro et al. (1997) analysed mouse A9 hybrid cell lines containing a single normal human chromosome 15, either paternal or maternal. Their findings indicated that all three $GABA_A$ receptor genes (in contrast to the findings in the mouse) were expressed only from the paternal allele. However, opposite results were obtained by Gabriel et al. (1998), who similarly used somatic cell hybrids containing either a maternal or a paternal human chromosome. RT-PCR analysis indicated that all three $GABA_A$ receptor genes were non-imprinted. These conflicting results may in part be attributable to variability in the degree to which somatic cell hybrids maintain the imprinting patterns of the donor human chromosomes.

iv) Telomeric to the GABA receptor genes is the non-imprinted *P* gene, associated with recessive tyrosinase-positive albinism (OCA2) (Spritz et al., 1997). Its heterozygous deletion is thought to underlie the mild hypopigmentation described in patients with PWS and AS type Ia (Table 1.2) (Saitoh et al., 1994, Cassidy, 1997, Jiang et al., 1999).

v) Low-copy repeat elements have been identified at both ends of the PWS/AS deletion interval, near to the deletion breakpoint hotspots on 15q11-q13 (Figures 1.3 and 1.4, BP-2 and BP-3). These represent members of one gene family, with an ancestral locus, *HERC2* on the telomeric side of the interval, and several duplicated copies. This repeat element (termed *D15F37*) is widely expressed, with a predominant transcript of 6-7 kb and a minor one of 15 kb. The ~15 kb transcript was termed *HERC2* (for Hect domain and RCC1 domain protein 2). It encodes a large predicted protein of 4834 aa and 528 kDa (Ji et al., 1999), that may function as a guanine nucleotide exchange factor and as an E3 ubiquitin ligase, potentially involved in the intracellular vesicular transport and degradation pathway. Evolutionary duplications of *HERC2* have lead to a family of both adjacent (on chromosome 15q13) and dispersed (15q11 and 16p11.2) copies. In the mouse, *Herc2* mutations result in neuromuscular secretory vesicle and sperm acrosome defects and juvenile lethality (Ji et al., 1999).

1.1.8.8 Molecular characteristics of the PWS/AS deletion on human 15q11-q13

As already described, the ~4 Mb PWS/AS deletion spans the region between *D15S1035* and *D15S165* (Christian et al., 1995). Two breakpoint clusters have been mapped centromeric to *ZNF127* (Figure 1.4), while the telomeric breakpoint is distal to the *P* gene (Mann and Bartolomei, 1999), within the *HERC2*-homologous repeat cluster. A fourth translocation cluster region in the 3′ region of *SNRPN* was mentioned above.

Figure 1.4. Map of the PWS/AS deletion region of human chromosome 15q11-q13 (not to scale). The map indicates the directions of centromere (cen) and telomere (tel), and the allelic expression patterns of inidividual genes are colour-coded as before (Figure 1.1). The open block arrow indicates the extent of the region frequently deleted in the PWS and AS, delimited by the markers *D15S1035* and *D15S165*. The two proximal and one distal common deletion breakpoints regions (BP-2 to BP-3) for PWS and AS deletions are indicated by zigzag lines, as is the translocation breakpoint cluster 3′ to *SNRPN*. The positions of duplicated sequences related to *HERC2* are indicated by grey ovals (Modified from Ji et al., 1999, Lee and Wevrick, 2000, De los Santos et al., 2000, Nicholls, 2000).

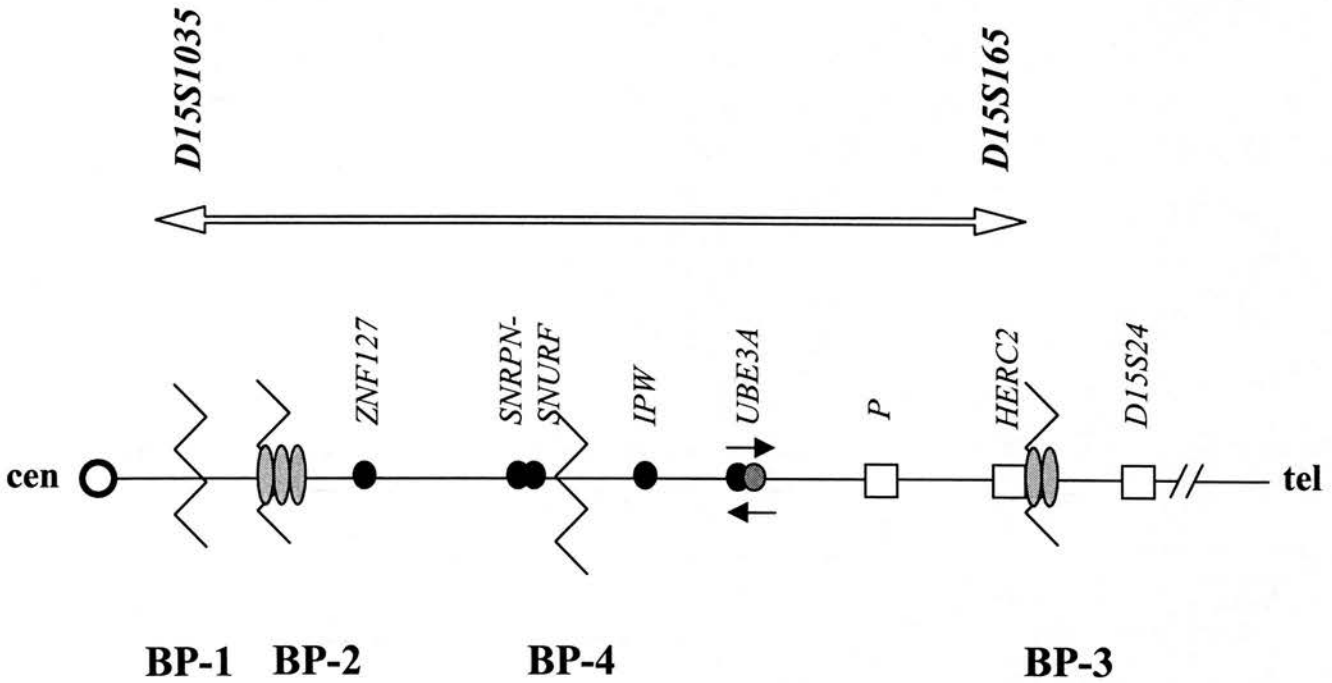


Figure 1.4. Map of human chromosome 15q11-q13 (not to scale) showing the frequent deletion breakpoints in this region

The 15q11-q13 region is involved in an unusually high number and diversity of recurring rearrangements. Besides the ~4 Mb PWS/AS deletion, over half of all supernumerary marker chromosomes result from rearrangements at the proximal and distal ends of 15q11-q13 (Christian et al., 1995). Other rearrangements involving this region include duplications, triplications and balanced reciprocal translocations (Christian et al., 1998). Considering the variety and frequency of these rearrangements, 15q11-q13 can be regarded as a "hot spot" for chromosome anomalies (Eichler, 1998).

Christian et al. (1995, 1998 and 1999) were able to identify four recurring breakpoints in 15q11-q13 in PWS and AS patients, numbered 1 to 4 (BP1-4). An analysis of 53 PWS and 33 AS deletion patients showed 44% of them to be deleted for the molecular marker *D15S542* ("Class I" patients), while the remaining 56% were not deleted for this marker ("Class II" patients). The centromeric breakage point for Class I patients is referred to as BP1 and that for Class II as BP2 (Figure 1.4) (Christian et al., 1999).

BP3, at the telomeric end of the deletion, lies within the *HERC2* repeat cluster; two duplicated copies of a 400 kb genomic region are found in tandem at this point. The last breakpoint region, BP4, has been located between markers *D15S1031* and *D15S1010* and has been described in cases of inv/dup, duplication and triplication of chromosome15 (Christian et al., 1999).

1.1.8.9 Origin of UPD in PWS and AS

As mentioned above, UPD accounts for 25% of PWS and 2-5% of AS. The incidence figures quoted above for the two disorders therefore imply frequencies of maternal and paternal UPD15 as respectively 1/40,000-1/60,000 and 1 in 400,000 to 1/1,000,000 respectively (see also Robinson et al., 2000). The more than tenfold difference between the frequency of matUPD and patUPD is attributed to the higher rate of nondisjunction in female gametogenesis. Using recombination analysis with microsatellites covering the entire chromosome 15, Fridman and Koiffmann (2000)

determined that heterodisomic PWS cases had originated by MI nondisjunction, while the cases of isodisomic PWS UPD15 had originated by postzygotic duplication. In five cases of AS with paternal UPD15, four isodisomies were detected. Three of these showed homozygosity for all markers, corresponding to a mitotic error, while the remaining case was defined as originating from a MII ND event. These results support the idea that most maternal UPD15 results from MI errors, with rare cases being MII or due to postzygotic errors, while on the other hand most paternal UPD15 is postzygotic in origin.

In 24% of the cases of matUPD15 examined by Robinson et al. (2000), the patients also showed an extremely skewed pattern of X-chromosome inactivation. As only those cases of trisomic zygote rescue with a relatively late origin of the diploid cell line will result in skewed X inactivation, this probably indicates a post-zygotic loss of the paternal chromosome in these cases.

1.1.8.10 The imprinting centre (IC) region of chromosome 15q11-q13

In addition to the typical ~4 Mb PWS/AS deletion discussed above, much smaller microdeletions (and other chromosomal rearrangements) upstream of *SNRPN* have been described in a group of patients that show evidence of abnormal imprinting in 15q11-q13. Careful molecular analysis has gradually enabled the narrowing down of the size of the shortest chromosomal region involved in these cases of PWS and AS (Woodage et al., 1994, Greger et al., 1994, Michaelis et al., 1995, Burke et al., 1996, Buiting et al., 1999).

Sutcliffe et al. (1994) described deletions of 5 kb and 60 kb in a familial and a sporadic case of PWS respectively. The deletions removed exon 1 of *SNRPN*, which is embedded in a differentially methylated CpG island. What was particularly interesting, however, was that these mutations caused altered methylation and loss of paternal gene expression over hundreds of kb. It was also noted that the 5 kb deletion was benign when maternally inherited. Based on these observations, the authors

proposed the presence of an imprinting control region or imprinting centre (IC) near exon 1 of *SNRPN*.

Further evidence for the existence and function of the IC region was described by Buiting et al. (1995). These authors demonstrated that different inherited microdeletions of the region were capable of causing both AS and PWS syndromes, and that rather than two completely separate paternal and maternal IC elements, there was a single *cis*-acting IC that regulated the chromatin structure, DNA methylation and gene expression through the domain (Dittrich et al., 1996, Buiting et al., 1999). Fine analysis, however, did enable the demonstration of adjacent, largely non-overlapping minimal regions of the IC that were deleted specifically in AS or in PWS cases (see below).

The disruption of the epigenetic program by the IC deletions means that while chromosome 15 exhibits a normal biparental mode of inheritance, AS patients have two chromosomes with a paternal identity (hypomethylation and biallelic expression of paternally expressed genes) while conversely, PWS patients have two chromosomes with a maternal identity (hypermethylation and silent paternal genes). In a sense, therefore, these patients have functional UPD, despite the biparental origin of their chromosomes (Horsthemke et al., 1997, Mann and Bartolomei, 1999).

The exact locations of the IC deletions in AS and PWS are different: AS deletions involve the upstream *BD* exons, while in PWS the deletions map slightly further 3′, encompassing *SNRPN* exon 1 (exon α or –1) (Dittrich et al., 1996, Saitoh et al., 1996). From analysis of two families with IC microdeletions, Ohta et al. (1999a) demonstrated that the AS shortest region of deletion overlap (AS-SRO) was ~35 kb centromeric to exon 1 of *SNRPN*. Buiting et al. (1999) similarly identified an AS family with a 5 kb IC microdeletion (the smallest found so far in AS) that further narrows down the AS-SRO to a 880 bp fragment containing two of the *BD* upstream exons. The analysis of the *BD* exons' structure has also indicated that the elements of the IC region have undergone multiple duplication events. Since the duplications

include not only the exons but flanking DNA as well, they are probably the result of chromosomal events such as unequal crossovers (Farber et al., 1999).

In the case of PWS it is the distal part of the IC that is deleted. The critical region is <4.3 kb spanning the *SNRPN* gene CpG island and exon 1, again as determined by the shortest region of overlap of microdeletions in PWS (PWS-SRO) (Dittrich et al., 1996, Ohta et al., 1999b).

The function of the IC is as yet not completely understood, and several models have been proposed to explain its action. In the model proposed by Dittrich et al. (1996), the IC consists of an imprintor and an imprint switch initiation site. The imprintor encodes the *BD* transcript; it is transcribed from the paternal chromosome only and acts in *cis* on the switch initiation site (which may be the *SNRPN* promoter, exon 1 or a site close by), possibly by introducing a local change in the chromatin structure. In the XX germline, a *trans*-acting factor specific for female germline cells is involved in completing the switch and/or initiating the bi-directional spreading of the maternal imprint on the paternally derived chromosome which switches off the imprintor activity.

In the absence of this maternal *trans*-acting factor (XY germline), the maternal imprint on the maternally-derived chromosome is lost, starting from the switch initiation site (spreading of the paternal imprint) (Dittrich et al., 1996, Ferguson-Smith, 1996, Kelsey and Reik, 1997). The model implies that in a given AS family, the grandfather is able to switch the imprint of his deletion-bearing maternal chromosome, because the maternal-to-paternal switch does not involve the activity of the *BD* exons. On the other hand, his daughter (the mother of the AS child) cannot switch the epigenotype, because the paternal-to-maternal switch depends on *BD* activity. Exon 1 (*SNRPN*) meanwhile, is required for establishing the maternal-to-paternal switching (Dittrich et al., 1996, Ferguson-Smith, 1996).

The suggestion that the IC controls imprint switching in the male and female germ line implies that IC microdeletions can transmit silently for multiple generations, but

that when the sex of the transmitting individual changes, the mutation blocks the appropriate resetting of the imprint specific for that individual germ-line. Thus, PWS IC mutations block the normal maternal-to-paternal imprint switch in the male germ line and AS IC mutations block the paternal-to-maternal imprint switch in the female germ line (Ohta et al., 1999b).

In a recent description of a family with PWS in male and female first cousins, Ming et al. (2000) demonstrated a submicroscopic deletion of *SNRPN.* In this family, the two cousins' fathers and two paternal aunts had the same deletion and were clinically normal. Since the surviving paternal grandfather did not carry the deletion, it was assumed to derive from the grandmother. Thus, an imprinting centre mutation can yield a "grandmatrilineal" inheritance pattern, in which a woman carrying the mutation (which affects imprinting when paternally transmitted) is at risk of having affected grandchildren through her unaffected sons. PWS does not become evident as long as the deletion is passed through the matrilineal line (Ming et al., 2000).

An alternative "two-step" model for imprint switching involves the erasure of the imprints, which requires the *SNRPN* CpG island and occurs in both germlines, followed by an active resetting. In this model, *BD* activity is involved specifically in resetting the maternal epigenotype in the female germline, where it is expressed from both "erased" chromosomes. In the absence of *BD* transcription in the male germline, a paternal imprint is acquired (Ferguson-Smith, 1996, Kelsey and Reik, 1997).

More recently, however, evidence has emerged that suggests that the PWS IC may act not just to "switch" epigenotype to paternal in spermatogenesis, but to maintain the imprint in somatic cells. Bielinska et al. (2000) described a PWS family in which the father is mosaic for an IC deletion of ~200 kb spanning *SNRPN* on his paternal chromosome. They found that the deletion chromosome had acquired a maternal methylation imprint in his somatic cells. The findings were supported by a chimaeric mouse model of the deletion, in which the chromosome with the paternal deletion has also adopted a maternal methylation pattern in somatic tissues. These results indicate that a paternal chromosome can postzygotically acquire a maternal imprint,

and therefore that the PWS IC element is not only required for establishing the paternal imprint in the paternal germ line but also for the postzygotic maintenance of this imprint. In this PWS family, unlike that of Ming et al. (2000), the IC deletion was inherited from the grandpaternal rather than the grandmaternal chromosome via the father. This indicated for the first time that contrary to the prediction of the "switch" models, the grandpaternal chromosome is not immune to rearrangements that ultimately result in PWS (Mann and Bartolomei, 1999).

Some experiments have been performed to address the possible *cis*-regulatory properties of the IC. In a transgenic *Drosophila* model, Lyko et al. (1997) demonstrated that a fragment from the region around *SNRPN* exon 1 can function as a parent-of-origin-independent transcriptional silencer. The DNA fragment used comprised nt -203 to +538 (relative to nt 1 of *SNRPN* exon 1) and the authors were able to narrow down the silencer activity to a 215 bp sequence (nt –203 to +12) that contained the *SNRPN* promoter. However, it is still unknown whether this result reflects the presence of a transcriptional silencer that is also active in mammalian cells.

Nuclear matrix binding assays have been used to demonstrate the presence of matrix attachment regions (MARs) ar high density within the IC region of 15q11-q13 (Greally et al., 1999). MARs are regions of DNA identified by their preferential association with the nuclear matrix (the structural "scaffold" that remains after treatments that deplete nuclei of most of their DNA) (Cockerill and Garrard, 1986). MARs are frequently found in association with regulatory elements such as enhancers and represors of transcription, origins of DNA replication, and *cis*-acting regulators of chromain structure and cytosine methylation (Rollini et al., 1999). In the PWS/AS region, MARs could exert a chromatin-organizing effect that proceeds differently on passage through the male and female germ line. The resulting allelic differences in chromatin structure dependent gene regulation would be recognized as genomic imprinting. According to this model, the heterochromatin-forming propensity of the IC must be differently modulated by other factors during male and female gametogenesis (Greally et al., 1999).

1.1.8.11 Histone acetylation in 15q11-q13

Chromatin immunoprecipitation revealed that the unmethylated CpG island of the active, paternal allele of *SNURF-SNRPN* was associated with acetylated histones, whereas the methylated maternally derived, inactive allele was hypoacetylated. When cells were treated with a DNA methyltransferase inhibitor, inducing demethylation of the CpG island, gene expression from the maternal allele was reactivated (Saitoh and Wada, 2000). These studies showed that *SNURF-SNRPN* is associated with parent of origin specific histone acetylation which is confined to the CpG island. The findings re-emphasize the intimate link between methylation and histone acetylation status in the epigenetic control of gene expression. It was also pointed out that since these experiments demonstrated that a gene that is silenced in PWS can be reactivated by drug treatment, there may be future possibilities for pharmacological therapy in disorders of genomic imprinting.

1.1.8.12 Asynchronous replication in 15q11-q13

By FISH analysis, the normal human chromosome 15 homologues were demonstrated to replicate asynchronously and to exhibit preferential association with each other during late S phase. Knoll et al. (1994) observed asynchronous replication in a high percentage of cells (25-40%) both from normal individuals and patients with PWS and AS with chromosome 15 deletions. In contrast, asynchronous replication was not found in PWS patients with maternal UPD. At least three different replication patterns were discernible at the telomeric end of the PWS/AS region: the paternal *D15S63* allele showed early replication, while *GABRA5* behaved oppositely (maternal early). The *P* gene replicated asynchronously but this occurs randomly with respect to its parental origin.

Further studies of a ~650 kb region that includes *SNRPN* were carried out by Gunaratne et al. (1995). This domain replicates late in lymphocytes but predominantly early in neuroblasts, where replication asynchrony was noted. A familial PWS deletion that encompasses part of the 5′ end of *SNRPN* resulted in the

loss of late replication of the domain in lymphocytes, when the deleted chromosome 15 is inherited paternally. Therefore, a potential allele-specific replication timing control region colocalises with the ICR. Furthermore, this region seems to control the replication of the paternal copy of the PWS domain in *cis*, and has no effect on the replication of a paternal allele in *trans* or on the replication of the maternal copy in *cis* or *trans*. This implies the existence of two different regions that control the replication of the paternal and maternal alleles of the PWS region.

1.1.8.13 Nuclease hypersensitivity of the *SNRPN* transcription unit

A region of 150 kb including *SNRPN* was examined for nuclease hypersensitivity in lymphoblastoid cell lines from PWS and AS individuals. This identified regions of hypersensitivity that were preferentially or exclusively parental-origin-specific (Schweizer et al., 1999): exon 1 of *SNRPN*, for example, is flanked by two hypersensitive sites on the paternal allele, one (pHS1) just upstream of exon 1 and the other (pHS2) ~1.7 kb downstream of exon 1 within *SNRPN* intron 1. Some other hypersensitive regions were identified on the maternal allele, one of which coincides with the AS minimal IC region and another of which lies within *SNRPN* intron 1 immediately downstream of the paternal-specific hypersensitivity site. In general terms, this study indicated a pattern whereby nuclease hypersensitivity of one allele correlated with hypermethylation on the allele contributed by the other parent. (Schweizer et al., 1999).

1.1.8.14 The mouse model for PWS and AS

The first candidate mouse models for PWS and AS were described by Cattanach et al. (1992 and 1997). Intercrosses between strains with chromosomal translocations were used to derive progeny with UPD of the PWS/AS-homologous region of chromosome 7. The maternal duplication (ie chromosomally comparable to PWS) caused a murine imprinting defect that was expressed as an early postnatal lethality, possibly resulting from impaired suckling activity. However, the paternal duplication

was not associated with any detectable phenotypic effect which might correlate to AS.

When targetted deletions within the region were made, an intragenic deletion within *Snrpn* resulted in no phenotypic abnormality, suggesting that mutations of *SNRPN* alone might not suffice to produce PWS. However, mice with a larger deletion, involving both *Snrpn* and the putative PWS-IC, lacked expression of *Zfp127, Ndn* and *Ipw*. Since these effects of the targetted IC mutation were similar to those of the human IC mutations, the role of IC in controlling imprinting appears to be conserved between humans and mice (Yang et al., 1998).

Gerard et al. (1999) generated *Ndn* mutant mice, and demonstrated that a paternally-inherited deletion allele resulted in post-natal cyanosis, dyspnoea, and lethality (within 30 hours). The lethality associated with the paternal *Ndn* deletion appeared to be due to a respiratory defect. It was therefore suggested that this is a relevant model for the neonatal respiratory distress (resulting from muscular hypotonia) often observed in humans with PWS. Necdin may thus be the first single gene within the PWS deletion interval for which a loss of function mimics some phenotypic aspects of PWS.

A null mutation in *Ube3a*, the murine orthologue of the AS gene, has also been produced by gene targetting. The maternally-deficient mice presented motor dysfunction, inducible seizures, and a defect in contextual learning and hippocampal long-term potentiation (LTP). LTP is an electrophysiological phenomenon whereby stimulation of presynaptic axons increases the strength of connections to postsynaptic neurons for days to weeks; it is regarded as a form of neuronal plasticity and is considered the strongest candidate cellular mechanism for learning and memory. This defect observed in the AS mouse model represented the first evidence for a role of ubiquitination in mammalian LTP (Albrecht et al., 1997, Jiang YH et al., 1998).

1.1.9 THE CLUSTER OF IMPRINTED GENES IN 11p15.5

The second imprinted gene cluster that has been a major focus for study in both human and mouse maps to human chromosome 11p15.5. Genetic pathology in this region is associated with Beckwith-Wiedemann syndrome (BWS). The region shows conserved synteny with a cluster of imprinted genes on distal chromosome 7 in the mouse (Hoovers et al., 1995, Yuan et al., 1996, Zubair et al., 1997, Ainscough et al., 1998, Paulsen et al., 1998) (Figure 1.5).

Figure 1.5. Simplified map of genes within the imprinted cluster on human chromosome 11p15.5 (not to scale). Individual genes mapping to this region are indicated by dark circles (paternally expressed), grey circles (maternally expressed) or open rectangles (non-imprinted). The small arrows show overlapping and antisense transcription. The two putative regional imprinting control centres are marked as IC. The three loci that have been specifically implicated in the pathogenesis of Beckwith-Wiedemann syndrome are indicated as BWS (modified from Hoovers et al., 1995, Alders et al., 2000 and Nicholls, 2000).

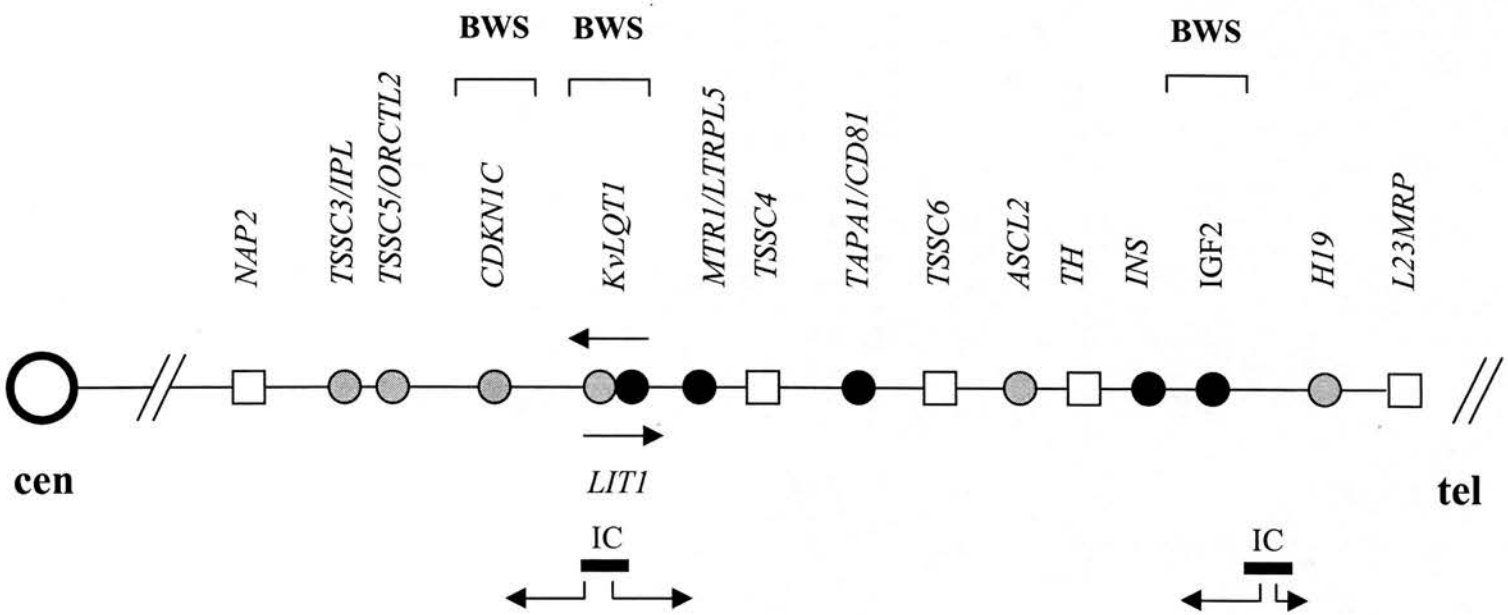


Figure 1.5. Schematic diagram of the locations of genes within the imprinted cluster on human chromosome 11p15

1.1.9.1 The Beckwith-Wiedemann syndrome

This syndrome (OMIM #130650) was described independently by Beckwith in 1963 and Wiedemann in 1964. It has a frequency of 1 in 14,000 births and is characterised by pre and postnatal overgrowth, macroglossia and anterior abdominal wall defects, typically omphalocoel (the EMG triad) (Weng et al., 1995). Variable features of BWS include organomegaly, transient hypoglycaemia, hemihypertrophy, genitourinary abnormalities and in about 5% of affected children, embryonal tumours (mainly Wilms' tumour but less frequently also adrenocortical carcinoma, hepatoblastoma or rhabdomyosarcoma) (Mannens et al., 1994).

The pathogenesis of BWS, like that of PWS/AS, is heterogeneous. About 15% of cases are familial, these displaying a maternal inheritance pattern. Balanced translocations involving 11p15 are seen in 1-2 % of BWS patients. Most of the remaining cases are sporadic (20% of these being associated with segmental/mosaic paternal UPD) (Weng et al., 1995). Inactivating mutations of the *CDKN1C* gene (encoding $p57^{KIP2}$) have been reported in 40% of the familial and 5% of sporadic BWS cases, and loss of the maternal imprinting of *IGF2* has been described in 30% of BWS patients (Reik and Maher, 1997, Hatada et al., 1996, Algar et al., 2000).

Several different translocation breakpoint cluster regions (BWSCR) have been defined in patients with BWS. BWSCR1 comprises 5 breakpoints clustered centromeric to *IGF2* and *H19*, that disrupt the gene *KVLQT1*. As discussed below, these BWSCR1 translocations have been postulated to cause BWS through indirect effects on the expression of one or both of the genes *CDKN1C* and *IGF2*.

BWSCR2 and BWSCR3, however, map distantly to the BWS imprinted gene cluster, lying respectively 5 Mb and 7 Mb proximal to BWSCR1 (Figure 1.6). BWSCR2 is defined by the breakpoints in two patients, and has been associated specifically with hemihypertrophy, a clinical finding not seen in patients with translocations involving BWSCR1 or BWSCR3 (Hoovers et al., 1995 Alders et al., 2000). The BWSCR2 translocations disrupt *ZNF215*, a tissue-specifically maternally expressed zinc finger

gene, but the precise nature of the link between this remote translocation site and the more telomeric BWS imprinted gene cluster remains unclear.

Figure 1.6. Locations of the BWS breakpoint regions on human chromosome 11p15 (map not to scale). Four of the genes in the region are indicated by the coloured rectangles as before; the directions of the centromere (cen) and telomere (tel) are also indicated. The three breakpoint regions are indicated as BWSCR1, BWSCR2, BWSCR3, the last two mapping 5 Mb and 7 Mb from BWSCR1 (modified from Hoovers et al., 1995, Alders et al., 2000 and Nicholls, 2000).

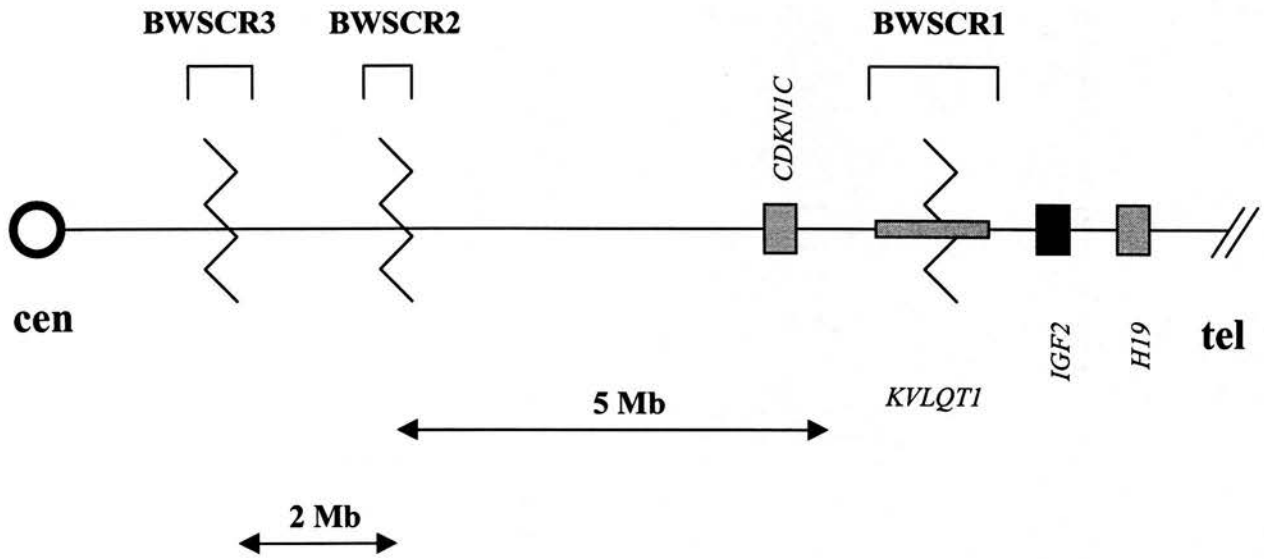


Figure 1.6 Localisation of the breakpoint regions on human chromosome 11p15

As in the PWS-AS region, imprinted and non-imprinted genes are interspersed within the BWS region. Specific features that have contributed to our understanding of imprinting are described below, taking the individual genes in a centromeric to telomeric order.

i) The *NAP2* gene (nucleosome assembly protein 2) encodes a highly acidic 375 aa protein, homologous to the yeast nucleosome assembly protein (NAP1) (Hu et al., 1996). *NAP2* is biallelically expressed. It has therefore been considered as defining the centromeric border of the distal BWS imprinted domain. However, the recent finding of an imprinted gene at BWSCR2, 5 Mb farther centromeric (Alders et al., 2000), suggests that this may underestimate the true extent of imprinting effects on 11p15.5.

ii) *TSSC3* (or *IPL*, *i*mprinted in *p*lacenta and *l*iver) was the first apoptosis-related gene demonstrated to be imprinted. It is maternally expressed in most tissues during human development (exceptions being the heart, testis and kidney). In adult brain and blood, *TSSC3* is biallelically expressed; however, in most neuroblastoma, medulloblastoma and glioblastoma tumours, it was found to be monoallelically expressed (Qian, et al., 1997, Muller et al., 2000). Given that *TSSC3* may have a growth inhibitory effect, it has been argued that suppression of its expression, due to retention or regaining of imprinting, is involved in the process of tumour development (Muller et al., 2000).

iii) The *ORCTL2* gene (*o*rganic *c*ation *t*ransporter-*l*ike 2) is also referred to as *IMPT1* and *TSSC5*. It has 2 transcripts, encoding predicted proteins of 253 and 424 amino acids. Although this gene is imprinted, its imprinting has been described as "leaky", meaning that the paternal allele is expressed at 20-30% the level of the preferentially expressed maternal allele and that there is little or no allelic bias at all in fetal fibroblast and lymphoblast cell lines. "Leaky" imprinting may reflect cellular heterogeneity, with certain populations of cells having complete repression of one allele and others showing biallelic expression (Cooper et al., 1998, Dao et al., 1998).

iv) The *CDKN1C* gene (or $p57^{KIP2}$) encodes a cyclin-dependent kinase inhibitor that induces G1 cell cycle arrest and is homologous to $p21^{CIP1/WAF1}$, a mediator of p53-directed cell cycle arrest. *CDKN1C* has a transcribed polymorphic hexanucleotide repeat within exon 1, that encodes a proline alanine repeat (PAPA). This gene is preferentially (90-95%) expressed from the maternal allele, except in the brain, where comparable expression of both alleles has been demonstrated (Hatada et al., 1995, Matsuoka et al., 1995 and 1996, Zhang et al., 1997, Algar et al., 2000).

A null mutation of *Cdkn1c* in the mouse causes umbilical abnormalities (omphalocoel) and respiratory problems (Zhang et al., 1997). Mutations of *CDKN1C* are also found in a proportion of BWS patients, most commonly in familial cases (Lam et al., 1999). The mutations in these cases are derived from unaffected carrier mothers, suggesting that loss of *CDKN1C* function results in BWS when the expressed allele encodes an inactive product and the other allele is repressed by genomic imprinting (Hatada et al., 1996).

v) Telomeric to *CDKN1C* is the *KVLQT1* gene, that encodes a voltage-gated potassium channel. This gene is maternally expressed in all human tissues except the heart. *KVLQT1* has 14 exons and spans a 300 kb region (Lee et al., 1997b, Mannens and Wilde, 1997).

KVLQT1 mutations cause long QT syndrome (LQT type Romano Ward), a dominantly inherited cardiac arrhythmia (Mannens and Wilde, 1997, Caspary et al., 1999). Deletions and insertions within *KVLQT1* have also been described in patients with the Jervell and Lange-Nielsen cardioauditory syndrome (JLN), an autosomal recessive condition characterised by congenital bilateral deafness associated with QT prolongation on the electrocardiogram, ventricular arrythmias and high risk of sudden death (Neyroud et al., 1997, Caspary et al., 1999, Mannens and Wilde, 1997).

The murine *Kvlqt1* gene is strongly expressed in heart, lung, gut, kidney and uterus. While its expression is maternal in early development, the paternal allele becomes increasingly active during development and juvenile animals show completely

biallelic expression. In humans, the lack of parent-of-origin effect in the LQT and JLN syndromes is due to the relative lack of imprinting in the heart, the main affected tissue (Gould and Pfeiffer, 1998, Caspary et al., 1998). Interestingly, although *KVLQT1* is involved in chromosomal rearrangements that are found in a small proportion of BWS patients, so far no BWS patients with LQT or JLN syndromes have been reported (Lee et al., 1997a).

vi) *KVLQT1* has an antisense transcript named *KVLQT1AS* or *LIT1* (long QT intronic transcript 1) that is paternally expressed but which in many patients with BWS is abnormally expressed from both the paternal and maternal alleles, (loss of imprinting, LOI). *LIT1* spans over 60 kb and is not translated (Lee et al., 1999, Mitsuya et al., 1999, Smilinich et al., 1999).

In the study of Lee et al. (1999a), 58% of BWS patients showed loss of maternal allele-specific methylation of the CpG island upstream of *LIT1*. The LOI of *LIT1* appeared not to be associated with LOI of *IGF2*, and LOI of *LIT1* was suggested to be the most common genetic alteration in BWS. (Also, the fact that *LIT1* and *IGF2* are independently affected in BWS patients has lent support to the view that in this region there are actually two imprinted domains (discussed in section 1.1.10.4) (Lee et al., 1999a). It has been suggested that the unmethylated *LIT1* CpG island acts as an insulator between *CDKN1C* and its enhancers, so that LOI of *LIT1* would result in suppression of *CDKN1C*. However, in mice lacking *Dmnt1*, where presumably both CpG islands are unmethylated, *Cdkn1c* remains active (John et al., 1999). Furthermore, some of the BWS patients with translocations within *KVLQT1* have been reported to show abnormal *IGF2* expression. Thus, there may not be complete independence between the regulation of the centromeric and telomeric regions of the 11p15.5 cluster (Horike et al., 2000).

vii) Telomeric to *KVLQT1* maps *MTR1* (named for resemblance to *MLS1*, a gene with homology to *melastatin*), also known as *LTRPC5* (large transient receptor potential-related channel 5). *MTR1* has 24 exons and spans 18.5 kb. It is alternatively spliced to encode two proteins of 872 and 1165 amino acids, both of which have

several predicted membrane-spanning domains. Transcripts have been identified in several fetal and adult tissues and also in Wilms' tumours and rhabdomyosarcomas. By RT-PCR analysis of somatic cell hybrids, exclusive expression of *MTR1* in cell lines carrying a paternal chromosome was demonstrated. However its murine orthologue (*Mtr1*) is biallelically expressed (Enklaar et al., 2000, Prawitt et al., 2000).

viii) The *TSSC4* and *TSSC6* genes were first identified on the so-called subchromosomal transferable fragment or STF, a 2.5 Mb fragment that suppressed *in vitro* growth of a rhabdomyosarcoma cell line. *TSSC4* and *TSSC6* code for 329 and 290 amino acid proteins respectively, with no close similarity to other known proteins. However *TSSC6* shows some weak similarity to *TAPA1/CD81* (located only 60 kb centromeric from *TSSC6*, between *TSSC4* and *TSSC6*) prompting the suggestion that these genes may have arisen by duplication followed by sequence divergence. *TSSC4* and *TSSC6* have a slight allelic expression preference in some tissues, but are mostly biallelically expressed (Lee et al., 1999b) in humans. In the mouse, there are some differences in expression, despite the conservation of the overall gene content and organization of the region. *Tssc4*, for example, unlike its human counterpart, is imprinted in mouse placenta (Paulsen et al. 2000).

ix) The *TAPA1/CD81* gene (see above) encodes a 26 kD membrane protein with four transmembrane domains, part of a signal transduction complex in T cells that respond to activated complement and has a critical role in early T-cell development. *TAPA1* belongs to the *TMSF* group of genes, and inactivation of some TMSF proteins has been shown to enable tumour progression. *TAPA1* is paternally expressed, but its mouse orthologue, although imprinted until e8.5, is biallelically expressed after this date (Reid et al., 1997).

x) The *ASCL2* gene (or *HASH2*) is maternally expressed only in the extravillus trophoblast. It is not expressed in benign androgenetic hydatidiform moles. However, expressio was detected in malignant moles, perhaps reflecting LOI in these tumours. In mice, mutation of the maternal *Mash2* orthologue is an embryonic lethal. This

may indicate that *HASH2*, also, is needed for normal development. If so, embryonic lethality of paternal UPD for 11p15.5 could provide an explanation for the fact that virtually all paternal UPD seen in BWS patients is mosaic (Alders et al., 1997).

xi) The imprinting status of the next gene, *TH* (tyrosine hydroxylase) has not been formally investigated. However, the recessive inheritance pattern of infantile Parkinsonism (Segawa disease, OMIM #605407) that results from *TH* mutations, suggests that *TH* is biallelically expressed.

xii) The imprinting of *INS* has been a matter of considerable interest. Giddings et al. (1994) studied the imprinting of *Ins2*, (the mouse homologous of *INS*) using single strand conformational polymorphism analysis. They showed that the gene was biallelically expressed in embryonic pancreas but that only the paternal allele was expressed in the yolk sac. The finding that *Ins2* was imprinted in mice supports observations that imply that the human *INS* gene may also be imprinted. Specifc alleles at a VNTR (variable number of tandem repeat) locus in the 5′ region of *INS* have been associated with an increased risk of insulin-dependent diabetes mellitus (IDDM), and there is a parental-origin dependence to this association (Bennett et al., 1997, Dunger et al., 1998). Conclusive evidence that *INS* is indeed also imprinted in humans has been recently described by Moore et al. (2001), who demonstrated that *INS*, like *Ins2*, is imprinted in the yolk sac.

xiii) The following two genes, *IGF2* and *H19*, have been intensively studied and used as a model for the regulation of genomic imprinting. They are therefore described separately in the next section (1.1.10.3).

xiv) The *LM23MRP* gene encodes an L23-mitochondrial related protein, an essential component of the large ribosomal subunit, which is predicted to contact and bind to ribosomal RNA. The gene is within 40 kb of *H19*, orientated "tail-to-tail", but is biallelically expressed. It has been considered to constitute the telomeric border of the imprinted 11p15.5 domain (Tsang et al., 1995).

1.1.9.2 The *IGF2* and *H19* genes

The *IGF2/Igf2* gene encodes a growth factor that seems to be involved mostly in fetal growth, and was one of the earliest identified imprinted genes. Mice with a targetted mutation disrupting *Igf2* displayed a heterozygous dwarfism phenotype, but only when the mutation was paternally inherited. The weights of mice with the mutation were ~60% of wild type (De Chiara et al., 1991). The inheritance pattern results from paternal monoallelic expression of *Igf2* during embryonic development (Giannoukakis et al., 1993, Ohlsson et al., 1993). The human *IGF2* gene is transcribed from four tissue-specific promoters, with transcription from the P1 promoter (preferentially active in adult liver) being biallelic (Vu and Hoffman, 1994, Vu et al., 2000).

H19 encodes a 2.5 kb RNA without open reading frame, and although its biological role is not entirely clear, it has been considered to have tumour suppressor activity (Zhang et al., 1993, Chen et al., 2000b). *H19* is transcribed only from the maternal allele and is very abundant in the developing mouse embryo, but after birth this gene is repressed in most tissues (Pachnis et al., 1988, Bartolomei et al., 1991, Zhang and Tycko, 1992).

Igf2 and *H19* lie ~ 90 kb apart, are reciprocally imprinted (as is also the case in humans) and have therefore been extensively studied in an effort to understand their regulatory interdependence (Thorvaldsen and Bartolomei, 2000). Regulatory regions important for the expression of both genes have been identified; these comprise differentially methylated regions (DMR) and a number of tissue-specific enhancers downstream of *H19* (Sasaki et al., 2000).

Three DMRs have been found in *Igf2*, two of which (DMR0 and DMR1) are located 5′ of the promoter P1 and one (DMR2) near the last exon (exon 6 in mice). In *H19*, there is a DMR ~2 kb upstream from the transcription start site, which is methylated in sperm and unmethylated in oocytes (Sasaki et al., 2000, Vu et al., 2000).

Leighton et al. (1995) identified a region essential for maintaining the regulation of both *H19* and *Igf2*. In this study, the regulation of the expression of *Igf2* and *Ins* was disrupted by a deletion spanning *H19* itself and 10 kb of the 5′ flanking sequence, suggesting a function of this *H19* region to imprint *Igf2*. By producing mice harbouring a 1.6 kb deletion (including the DMR) it was demonstrated that *H19* was activated and *Igf2* expression was reduced when the deletion was paternally inherited and that conversely *H19* expression was reduced and *Igf2* was activated upon maternal inheritance of the deletion.

It was deduced from such studies that the *H19* DMR (that spans ~2 kb and is 2 kb 5′ of *H19*) is required for the reciprocal imprinting of the two genes and also (to judge from the timing of its differential methylation) that it serves as the allelic mark distinguishing paternal and maternal alleles. This region has also been shown to act as a transcriptional silencer in *Drosophila* (Lyko et al., 1997, Ripoche et al., 1997, Thorvaldsen et al., 1998). In addition, by conditional deletion transgenic experiments, Srivastava et al. (2000) demonstrated that the presence of the *H19* DMR on the maternal chromosome is continually needed at different stages of development in mice in order to keep *Igf2* silenced.

Endoderm and muscle-specific enhancers downstream of *H19* are also involved in the regulation of *Igf2/H19* imprinting. By deleting the endoderm enhancers, Leighton et al. (1995) observed that neither *H19* nor *Igf2* was expressed on the chromosome with the deletion, indicating that both genes utilise the same enhancers, albeit on different parental chromosomes. It was also observed that when an extra set of enhancers was inserted midway between *Igf2* and *H19*, the normally silent maternal *Igf2* allele was activated in endodermal tissues without changing the level of *H19* expression, indicating that the endoderm enhancers in their correct location are essential for the imprinted expression of both *Igf2* and *H19*. At least five enhancer regions have been identified, and interestingly, they resemble a large cluster of tissue-specific enhancers at the locus control region (LCR) of the β globin gene cluster (Webber et al., 1998, Li Q et al., 1999, Sasaki et al., 2000).

Such observations led to the enhancer competition model, which was based on the similar patterns of expression of *Igf2* and *H19* in many endoderm and mesoderm derived tissues, implying the sharing of tissue-specific enhancers. The enhancers would preferentially activate *H19* on the maternal chromosome (perhaps due to proximity or to high affinity), leaving the maternal *Igf2* allele silent or only weakly expressed. On the paternal chromosome, the enhancers cannot interact with *H19* because the 5′ region of *H19* is heavily methylated and assumes a closed chromatin structure; thus the enhancers would engage with *Igf2* (Bartolomei and Tilghman, 1997, Sasaki et al., 2000).

This model suggests that *H19* transcription on the maternal chromosome is required to maintain the silence of the maternal *Igf2* allele. To test this prediction, Schmidt et al. (1999) examined a transgenic model in which the *H19* promoter and structural gene had been deleted. However, the absence of transcription from the *H19* promoter had no effect on the imprinting of *Igf2* in liver, and only a small effect in skeletal muscle. This led to a modified version of the enhancer-competition theory, in which it was proposed that *Igf2* imprinting is determined by a methylation-regulated chromatin boundary at the *H19* DMR. One experiment suggesting such a property for the *H19* DMR consisted of moving the DMR to a position downstream of the *H19* gene, between the endodermal and skeletal muscle enhancer elements. The DMR was thus positioned between between the *H19* promoter and the muscle enhancers. On maternal inheritance, activation of *H19* in skeletal muscle but not in liver was blocked, consistent with the presence of insulator activity being exerted by the ectopic *H19* DMR (Kaffer et al., 2000).

Two maternal-specific nuclease hypersensitive regions, HS1 and HS2, related to the DMR upstream of *H19*, constituted good candidate sites for the binding of proteins involved in the structure of the putative chromatin boundary (Hark and Tilghman, 1998, Ben-Porath and Cedar, 2000). Boundary elements are generally defined as *cis*-acting DNA sequences that delineate specialised gene expression domains in chromatin. Thus when placed at either side of a gene and its regulatory elements, they insulate that transcription unit from the regulatory elements of neighbouring

genes (Sasaki et al., 2000). Analysis of the *H19* HS1 and HS2 revealed that they overlap with several short CG-rich repetitive elements that are recognition sites for the zinc finger protein CTCF (Kanduri et al., 2000, Pfeifer, 2000, Thorvaldesen and Bartolomei, 2000). CTCF (CCCT-binding factor) is a transcription factor that plays a role in insulator activity at the chicken β globin locus (Bell and Felsenfeld, 2000). Its activity is sensitive to the methylation state of the DNA target site; zinc-fingers proteins such as CTCF prefer to interact with the major groove of DNA and presumably the projection of the methyl group attached to cytosine prevents the zinc finger from recognising key bases within the recognition sequence (Hark et al., 2000, Bell and Felsenfeld, 2000, Wolffe, 2000).

The model currently proposed to mediate the reciprocal imprinting of *Igf2* and *H19* is that the maternal chromosome represents the default or unimprinted state. CTCF binds to the insulator element between the two genes, and the insulator-CTCF complex blocks the *Igf2* gene from the enhancer that is positioned 3′of *H19*. Consequently, only the *H19* gene is active. When the DMR and the *H19* gene are methylated, as on the paternal chromosome, CTCF fails to bind to the insulator and then the 3′ enhancer can activate the *Igf2* promoter (Wolffe, 2000) (Figure 1.7).

Figure 1.7. Regulated insulator activity at the *Igf2-H19* locus.

A) The *Igf2* and *H19* genes are about 90 kb apart (not to scale). The small arrows indicate the start site and direction of transcription of each gene. The positions of the insulator (at the imprinting control region) and the enhancer are indicated. There is in fact more than one tissue-specific downstream enhancer region.

B) On the maternal unmethylated locus, CTCF can bind to the insulator and prevent the 3′ enhancer from activating the *Igf2* promoter. The enhancer can still efficiently activate the unmethylated *H19* promoter.

C) On the paternal methylated locus, CTCF can no longer bind to the insulator because the DNA is methylated (Me) and insulator activity is therefore lost. In addition, methylation of the H19 gene sequence represses transcription directly. Enhancer activity is now focused on activation of the *Igf2* promoter. (As in Wolffe, 2000.)

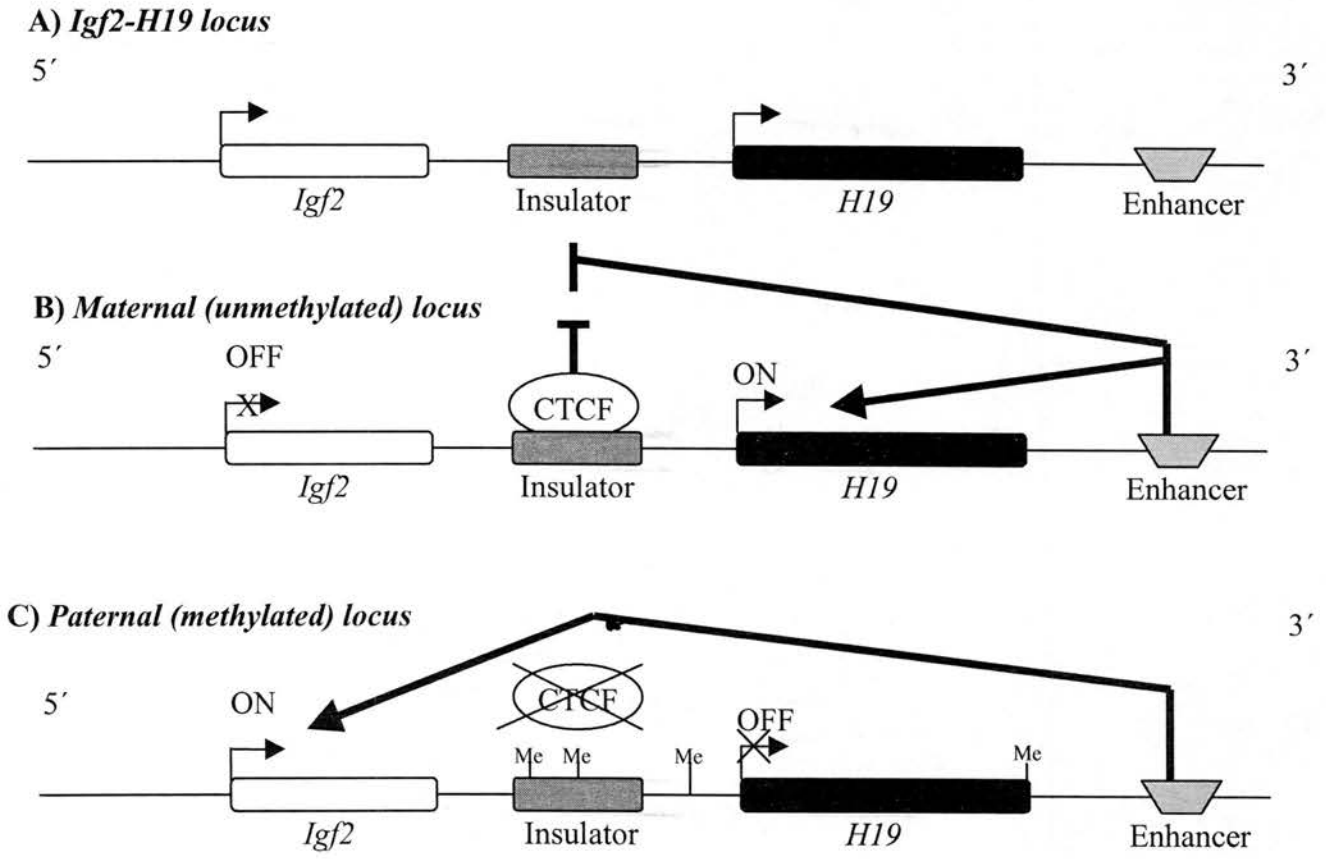


Figure 1.7. Regulated insulator activity at the *Igf2-H19* locus

1.1.9.3 The two domain hypothesis in Beckwith-Wiedemann syndrome

As described above, structural or regulatory alterations of several different genes have been implicated in the pathogenesis of BWS. The patterns of abnormality have led to the suggestion that there are two distinguishable imprinted domains within the 11p15 BWS locus. One of these involves the antisense transcript *LIT1*, and the second is related to *IGF2* (Horike et al., 2000). Despite their proximity and their joint involvement in BWS, there is so far little direct evidence for a mechanistic link between the imprinting of the two domains. For example, their timing and direction of imprinting in the germline differ; the *H19* paternal allele is active until late in spermatogenesis, while the *CDKN1C* maternal allele is silent and is activated by passage through the female germline late in oogenesis (John et al., 1999).

Mouse models have been used to investigate the role that each of the two imprinted domains may have in the development of BWS. A mouse double mutant for both *Cdkn1c* and *Igf2* exhibits aspects of BWS that are not seen in other models involving each gene independently, such as macroglossia and a marked placentomegaly (Eggenschwiler et al., 1997).

The fact that BWS can apparently result from either biallelic expression of *IGF2* or *CDKN1C* mutations, as well as the corresponding overlapping phenotypes in mice, strongly suggest that these genes may act in a common growth control pathway. Grandjean et al. (2000) demonstrated that *Cdkn1c* expression is down-regulated in mice with high serum levels of *Igf2,* suggesting that the effects of increased *IGF2* in BWS may be mediated in part through a decrease in *CDKN1C* expression. Interestingly, this effect was not observed in the placenta, where there is evidence suggesting that *Igf2* acts through a different (and currently unidentified) receptor.

Similarly, to investigate the mechanism by which mutations of the putative centromeric imprinting centre result in BWS, Horike et al. (2000) analysed human chromosomes carrying a targeted deletion of the *LIT1* CpG island in somatic cell hybrids. This deletion abolished *LIT1* expression on a paternal chromosome, and this

was accompanied by activation of the normally silent paternal allele of multiple imprinted genes in the centromeric BWS domain (including *KVLQT1* and *CDKN1C*). However, the deletion had no effect on the regulation of *H19*. These studies suggest that *LIT1* has a role in the BWS region that includes silencing of *CDKN1C*, thus reinforcing the idea that *LIT1* is an imprinting control element.

1.1.10 OTHER IMPRINTED LOCI IN HUMANS

Besides the two intensively-studied clusters of imprinted genes on chromosomes 11 and 15, other regions of the genome deserve attention because they illustrate specific points relevant to the regulation of imprinting. Two of these regions are on chromosome 6, and two on chromosome 7.

1.1.10.1 Imprinted genes on chromosome 6

The 6q24 chromosomal region has been associated with the rare disorder transient neonatal diabetes mellitus (TNMD). This rare disorder is characterised by intrauterine growth retardation, dehydration and failure to thrive. The diabetes results from a congenital lack of insulin secretion, and although recovery occurs by 18 months of age ~40% of patients relapse and develop permanent type 2 diabetes later in life (Temple et al., 2000, Gardner et al., 2000).

Two imprinted genes have recently been described in 6q24. These are *ZAC1* and *HYMAI. ZAC1* is paternally expressed, and encodes a zinc finger protein, which promotes apoptosis and cell cycle arrest. It is also a transcriptional regulator of the type 1 receptor for the pituitary adenylate cyclase activating peptide ($PACAP_1$-R), an intra-islet regulator of insulin secretion. *ZAC1* is thus a strong functional candidate for being involved in TNDM, and abnormal methylation of the CpG island associated with it has been demonstrated in some TNDM patients (Gardner et al., 2000, Kamiya et al., 2000). TNDM, therefore, may be added to the short list of human diseases (also including PWS, AS, BWS, and PHP-Ib, discussed later) that result from abnormal control of imprinting. *HYMAI* (hydatiform mole associated transcript) is located 70 kb upstream of the coding region of *ZAC1*, and associated

with the same differentially methylated CpG island that represents part of the *ZAC1* transcription unit (Arima et al., 2000). It is a non-coding transcript that like *ZAC1*, is paternally expressed, and may be considered part of the same transcription unit. The extent of the 6q24 imprinted domain is presently unknown.

IGF2R (or *M6PR*) maps in humans to chromosome 6q26, and codes for the insulin-like growth factor type 2 receptor, for which the ligand is IGF2 (Barlow et al., 1991, Bartolomei et al., 1997). IGF2R is involved in targeting of lysosomal enzymes from the Golgi apparatus to the lysosomes, as well as in the internalisation and degradation of IGF2 (O'Dell and Day, 1998, Jirtle, 1999). The corresponding *Igf2r* gene is imprinted in the mouse, being expressed from the maternal allele except in CNS where is biallelically expressed (Vu and Hoffman, 1994). Human *IGF2R* has been reported not to be imprinted, in that it is biallelically expressed (Treacy et al., 1996). However, the intragenic DMR2 (see below) does show maternal allele-specific methylation in humans (Smrzka et al., 1995). In fact, it has been suggested by others that there is tissue-specific monoallelic expression of *IGF2R*, and that this may also be a polymorphic trait (Xu et al., 1993, Treacy et al., 1996). Loss of heterozygosity (LOH) or abnormal allelic expression of *IGF2R* has been detected in some human cancers (Kong et al., 2000).

The murine *Igf2r* gene has two DMRs: DMR1, located at the promoter, and DMR2, an intronic CpG island. Detailed experimental analysis has defined DMR2 as a 113 bp box that includes an 8 bp *de novo* methylation signal (DNS) at its 3′ end and a 6 bp allele-discriminating signal (ADS) that prevents the paternal allele from undergoing *de novo* methylation (Stoger et al., 1993, Birger et al, 1999). As previously mentioned, this intragenic DMR is the promoter for an antisense RNA transcribed from the paternal allele. The antisense RNA (named *Air* for antisense *Igf2rRNA*) terminates more than ~100 kb distant, within the flanking non-imprinted gene *Mas1* gene (Smrzka et al., 1995, Wutz et al., 1997, Lyle et al., 2000). Despite the functional evidence (see section 1.1.8.6) that *Air* represses *Igf2r* transcription in *cis* on the paternal allele, there may be dissociation between *Air* and the regulation of *Igf2r* in some tissues. Thus, *Air* is derived only from the paternal allele in CNS,

where it is unmethylated while the methylated maternal allele is silent (Hu et al., 1999), even though *Igf2r* is biallelically expressed in this tissue.

1.1.10.2 Silver-Russell syndrome and imprinting on human chromosome 7

The Silver-Russell syndrome (SRS) is characterised by pre and postnatal growth failure and dysmorphic features. Though it may be etiologically heterogeneous, in 7% of SRS cases maternal UPD7 has been demonstrated, suggesting that at least one gene on this chromosome is imprinted (Preece et al., 1997, Bernard et al., 1999).

De novo duplication of the maternal chromosome 7p11.2-p13 has also been reported in SRS patients. This suggests that the intrauterine growth retardation seen in maternal UPD7 may result from an increased dose of a paternally imprinted gene on chromosome 7 rather than the absence of a paternal growth-promoting gene (Devriendt, 2000, Monk et al., 2000). The duplicated region contains the genes for growth factor receptor-binding protein 10 (*GRB10*) and insulin like growth factor-binding proteins 1 and 3. Paternal UPD of this region in mouse causes prenatal overgrowth, and in human lymphocytes the *GRB10* genomic interval has been shown to replicate asynchronously (Monk et al., 2000)

Yoshihashi et al. (2000) found that *GRB10* was monoallelically expressed in human fetal brain and was transcribed from the maternal allele in somatic cell hybrids. In two of 58 SRS patients, they found a maternally-derived P95S substitution. They suggested that these findings indicated that these maternally transmitted mutant alleles contributed to the SRS phenotype. However, it now appears that *GRB10* imprinting is very complex. Blagitko et al. (2000) demonstrated expression from both the paternal and maternal allele, in a highly tissue- and isoform-specific fashion. In human fetal brain, *GRB10* was paternally expressed, in contrast to the mouse in which it is maternally expressed. This may represent the first example of a gene that is oppositely imprinted in human and mouse. The role of *GRB10* in SRS remains uncertain.

There are also imprinted genes on human chromosome 7q. *MEST* (*PEG1*) is the orthologue of *Mest/Peg1*, the first imprinted gene mapped to mouse chromosome 6 (Kaneko-Ishino et al., 1995), and is expressed only from the paternal allele. The disruption of *Mest* by gene targeting in embryonic stem cells showed that the mutation was silent on passage through the female germ line. On paternal transmission, the mutant allele caused embryonic growth retardation and abnormal maternal behaviour (Lefebvre et al., 1998). Another gene in 7q32, *COPG2*, (coatomer protein complex subunit γ2) has been described as imprinted (Blagitko et al., 1999) but this finding is now in doubt because of overlap between transcripts derived from *COPG2* and *MEST* (Yamasaki et al., 2000).

1.1.11. IMPRINTED REGIONS IN MOUSE

Some clusters of imprinted genes have been identified through studies in the mouse, with subsequent extrapolation to humans. One example is the region of mouse chromosome 7 that contains *Peg3* (human *PEG3* is on 19q13.4). *Peg3* was identified during a subtractive cDNA screen between parthenogenetic and biparental mouse embryos (Kuroiwa et al., 1996), it encodes a paternally expressed zinc-finger protein. Human *PEG3* has also recently been shown to be paternally expressed (Murphy et al., 2001). Targetted mutation of *Peg3* results in a phenotype characterised by growth retardation and altered maternal behaviour that results in death of the offspring. This alteration has been attributed to deficient neuronal connectivity (Li L. et al., 1999, 2000).

Another paternally-expressed gene in the same region of mouse chromosome 7 is *Ocat* (ossification centre-associated transcript). Its exon 1 is only 200 bp upstream of *Peg3*, and *Ocat* is transcribed away from *Peg3* (Szeto et al., 2000). *Zim2* is a further zinc finger gene, located 25 kb downstream of *Peg3*. However, the *Peg3* and *Zim2* transcripts share identical 5′ends, composed of 7 small exons. Alternative splicing events connect these 7 exons either with the remaining 2 exons of *Peg3* or with the remaining 4 exons of *Zim2*. The two genes share identical transcription start sites and probably therefore also share upstream regulatory elements (Kim et al., 2000).

Another cluster of two oppositely imprinted genes, Delta-like (*Dlk1*) and *Gtl2* has been described on mouse chromosome 12. The two genes are 120 kb apart and are transcribed in the same orientation. The paternal-specific expression of *Dlk1*and maternal-specific expression of *Gtl2* appear to be closely coordinated in a fashion similar to that of the *Igf2/H19* pair (Schmidt et al., 2000). This reciprocally imprinted expression pattern is again conserved in the the *DLK1* and *GTL2* human orthologues on chromosome 14q32 (Wylie et al., 2000).

1.1.12 X- CHROMOSOME INACTIVATION

As described above, imprinted genes often share several general characteristics; they tend to be located in clusters, to be regulated by a *cis*-acting imprinting centre, and to be associated with antisense transcripts. These properties also feature in the process of X inactivation, which is an interesting model for investigating similarities between the two inactivation mechanisms.

X inactivation is a mechanism for equalising the sex chromosome gene dosage in males and females (Latham, 1996). It is believed to require several different steps, namely: *counting* the number of X chromosomes in a cell (n), randomly *choosing* (n-1) X chromosomes to be inactivated per diploid genome, and *initiating* inactivation of these chromosomes (Panning and Jaenisch, 1998, Pereira and Vasques, 2000). There is normally an equal probability that either X chromosome will be inactivated in a given cell (Chandra and Brown, 1975, Penny et al., 1996). In marsupials, however, the maternal X is preferentially active and the paternal X is inactive (Sharman, 1971).

X inactivation starts at the X-inactivation centre (Xic/XIC), defined as the minimum *cis*-acting region required for an X chromosome to be inactivated. In humans XIC maps to Xq13 (Brown et al., 1991). Two genes have been identified at Xic: *Xist* (for X-inactive specific transcript) and *Tsix*, a gene antisense to *Xist* (Penny et al., 1996, Lee and Lu, 1999). *Xist* is expressed exclusively from the inactive X chromosome. It spans ~40 kb and encodes a ~15 kb non-coding RNA transcript, which is located in

the nucleus and is required in *cis* to propagate X inactivation. Human X:autosome translocations containing XIC reveal spreading of this inactivation effect onto the autosomal segment (Clemson et al., 1996, Penny et al., 1996).

As for autosomal imprinted genes, methylation also participates in the regulation of Xic. In male somatic tissues *Xist* is hypermethylated at multiple CpGs at its 5′ and 3′ ends as well as some sites within the gene. In females, methylation of *Xist* is incomplete but consistent with methylation of one allele (the active X) per cell (Belmont, 1996).

The 5′ regions of many genes are methylated on the inactive X chromosome (Xi) and unmethylated on the active chromosome (Xa). Meanwhile, the 5′ regions of genes that escape X inactivation remain unmethylated on both alleles (Tinker and Brown, 1998). Several (though not all) of the latter are clustered in and around the pseudo-autosomal region and in Xp11, suggesting that X inactivation, like imprinting, is a regional phenomenon, whereby entire blocks of genes are coordinately regulated (Brown et al., 1997).

In the process of inactivation, the Xi acquires particular features that differentiate it from Xa: Xi is late replicating, highly methylated on CpG islands and rich in hypoacetylated histone H4, all markers of gene inactivity (Hansen et al., 1996, Pereira and Vasques, 2000). The silent *XIST* gene on Xa also replicates before the expressed allele on Xi, supporting the view that replication timing is determined by the large, multi-replicon domain in which a gene is located, and not necessarily by its expression state (Gartler et al., 1999).

As for some imprinted genes, an antisense transcript has been identified at the *Xist* locus. *Tsix* overlaps the 3′ end of *Xist* and is transcribed in the reverse orientation. Initially biallelically expressed, it becomes monoallelically expressed at the onset of X inactivation, marking only the future Xa (Lee and Lu, 1999). During differentiation, loss of *Tsix* transcription seems to correlate with the stabilisation of *Xist* RNA on the Xi. Thus, the level of *Xist* RNA before differentiation may be

controlled by *Tsix* through an antisense RNA-mediated mechanism (Pereira and Vasques, 2000), as also suggested for some imprinted autosomal genes. Upon differentiation, *Tsix* is repressed on the future Xi, leading to stabilisation of the allelic *Xist* RNA. At the end of differentiation *Tsix* is off in both Xs, but *Xist* is expressed only from Xi. Based on these results it has been proposed that the initiating element for X inactivation is a repressor of *Tsix* (repTsix) that is expressed during differentiation and X inactivation. This repressor starts the cascade of events leading to X inactivation by associating with the *Tsix* promoter in all but one X chromosome per diploid cell (Pereira and Vasques, 2000).

1.1.13 EPIGENETIC CHANGES IN CANCER

Disturbances of genomic imprinting have been implicated in the development of cancer. Also, some of the syndromes caused by abnormalities of imprinting confer a risk of particular tumour types. The Knudson "two hit" model for recessive oncogenesis has been modified to take account of the effects of imprinting. Classically, the two "hits" required for loss of tumour suppressor gene function are observed as deletions (manifesting, if acquired, as loss of heterozygosity, LOH) or point mutations (Jones and Laird, 1999, McBurney, 1999, Tycko, 2000). In the case of dominantly-transmitted cancer-predisposing syndromes, one of the "hits" is inherited. However, for an imprinted gene, the normal epigenetic silencing of one allele can be considered as another kind of first "hit". The second "hit" could occur through LOH or epigenetically, silencing the tumour suppressor gene through DNA methylation and histone acetylation (Feinberg, 1993, Jones and Laird, 1999, Tycko, 2000).

Loss of imprinting (LOI) of growth promoting genes is also involved in cancer. This refers to the failure of the cell to maintain normal monoallelic expression (Feinberg, 1999). LOI may thus result in increased expression of an imprinted proto-oncogene, Furthermore, mutational inactivation of an IC has the potential to alter the regulation of multiple tumour suppressor genes or proto-oncogenes (Jirtle et al., 1999).

LOI at the *IGF2* locus has already been referred to in Beckwith-Wiedemann syndrome. Biallelic expression of *IGF2* has also been detected in 70% of Wilms' tumours, in 40% of colorectal cancers (compared to 13% of samples of normal colon; Ogawa et al.1993, Feinberg, 1999) and in 30% of breast cancers (Wu et al., 1997). LOI of both *IGF2* and *H19* has been identified in cervical and in ovarian cancer (Chen et al., 2000b). In contrast, *IGF2* imprinting status has generally been maintained in infant leukaemias and neuroblastomas (Hattori et al., 2000). LOI of *IGF2/H19* is further discussed below.

In neuroblastomas, deletions in distal 1p36 are preferentially maternal in origin, indicating that this region may contain an imprinted tumour suppressor gene (Caron et al., 1993). One candidate gene mapping to this region is *TP73* (OMIM *601990). *TP73* is maternally expressed and encodes a protein (p73) very similar to p53. Occasional LOH of *TP73* has been observed in various tumours, but so far no convincing evidence for somatic mutations of the gene has been accumulated. Thus p73 may be much less important than its homologue p53 in respect of tumour suppressor activity. However, a role for *TP73* in the pathogenesis of neuroblastoma has not been totally excluded; it could contribute to this process through a gene-dosage effect or secondary to transcriptional silencing of the remaining allele (Hu et al., 2000, Judson et al., 2000).

Epigenetic silencing of an imprinted tumour suppressor gene could involve histone deacetylation and/or DNA methylation. Deregulation of histone acetyltransferase (HAT) and deacetylase (HDAC) activity is certainly frequent in human cancer (Pedone et al, 1999). Inhibition of HATs or inappropriate activation of HDACs could therefore be part of the mechanism causing loss of imprinting in human cancer. Abnormalities in methylation have been studied more intensively, however. The von Hippel-Lindau gene (*VHL*) is an example of a tumour suppressor that may be silenced by methylation. Germline *VHL* mutations predispose to renal cell carcinomas and hemangioblastomas, while somatic mutations have been found in sporadic renal cell carcinomas. Also, in ~20% of the latter, the *VHL* promoter is hypermethylated (Tycko, 2000). This therefore appears to be a frequent alternative to

LOH as a route to the somatic "second hit" . Other tumour suppressor genes that can be inactivated by hypermethylation include *hMLH1* (colonic, gastric and endometrial cancer) and *p16/INK4A* (involved in multiple types of cancer) (Tycko, 2000).

Abnormalities in tumour DNA methylation are also reflected in the activity of the methyltransferase enzymes. RNA and protein levels of DNMT3A, DNMT1 and DNMT3B are elevated in several tumours (Robertson et al., 1999, Tycko, 2000). Whether this can be causally related to tumour suppressor hypermethylation, though, is unclear. In a study of colon carcinoma cells in which exons 3, 4 and 5 of *DNMT1* were disrupted, it was found that DNMT activity was only decreased by ~20%, implying that other methylases may take over this activity (Rhee et al., 2000).

As mentioned, abnormalities in the imprinting of *H19* and *IGF2* have been frequently described in Wilms' tumour (WT) and other tumours. The biallelic silencing of *H19* in WT is accompanied by abnormal *H19* methylation in tumour and surrounding normal kidney. This suggested that an epigenetic error affecting the DMR of *H19* occurred early in embryogenesis (Frevel et al., 1999). Silencing of *H19* accompanies LOI of *IGF2* in Wilms' tumours, suggesting alterations of these genes' coordinated epigenetic regulation as the mechanism underlying the participation of this locus in tumorigenesis (Steenman et al., 1994).

LOH or UPD at an imprinted locus may result in the deletion of the only functional copy of an imprinted tumour suppressor gene (Jirtle et al., 1999). Alternatively, epigenetic changes may be capable of inactivating the expressed allele of an imprinted tumour suppressor gene. *H19* is one example of an imprinted gene that has been suggested to have tumour suppressor activity. Dao et al. (1999) demonstrated that in several childhood tumours the normal monoallelic methylation of *H19* and its upstream sequences has often been converted to an abnormal pattern of biallelic hypermethylation. As a result of this change both alleles resembled the imprinted paternal allele of *H19*.

Abnormal imprinting of genes in 11p15.5 has also been seen in a large proportion (51%) of hepatocarcinomas (Schwienbacher et al., 2000b). Loss of the normal maternal-specific methylation at the putative BWS IC2 was associated with silencing of the maternal allele of *CDKN1C* and *IGF2* in these tumours. The authors refer to this phenomenon as "gain of imprinting". From studies such as these, it is certainly clear that imprinted regions are susceptible to epigenetic alterations in their expression patterns during tumorigenesis, and that this susceptibility may result in the common involvement of imprinted regions in cancer.

The idea that epigenetic changes at imprinted loci can be involved in tumorigenesis has been extended further, to incorporate polymorphic variation in imprinting control as a cancer risk factor. As described above, imprinting of *M6P/IGF2R* in humans seems to be a polymorphic trait, with most individuals having biallelic expression. Given the evidence that *IGF2R* can act as a classical tumour suppressor gene, with LOH and "second-hit" point mutations in liver tumours (de Souza et al. 1995), Jirtle et al. (1999) have suggested that individuals who have an imprinted *M6P/IGF2R* gene may have an increased susceptibility to tumour development because they already have an epigenetic "first hit".

The Wilms' tumour gene *WT1* has also been intensively studied with respect to imprinting and cancer development. WT arises from an abnormal development of embryonal kidney cells. The abnormal regulation of 11p15.5 genes in WT has been alluded to above. A proportion of WT, though, results from mutations of the *WT1* gene, which is also required for normal kidney development (Ramani and Cowell, 1996). *WT1* has been found to be imprinted in 40% of preterm human placentas, and since its imprinted status does not correlate with gestational age, *WT1* imprinting is a polymorphic trait (Nishiwaki et al. 1997). In fibroblasts and lymphocytes, too, polymorphic *WT1* imprinting has been seen, though here the paternal rather than the maternal allele is expressed (Mitsuya et al. 1997). This finding, for a known tumour suppressor gene, again suggests, as for *IGF2R*, the possibility that such a polymorphic trait could influence both tissue- and individual-dependent

susceptibilities to cancer and possibly also the response to medication (Bunzel et al., 1998, Jirtle et al., 1999).

At the *WT1* locus, an antisense promoter (*WT1-AS*) corresponding to a DMR located in intron 1 has been suggested to be involved in control of WT1 expression. *WT1-AS* does not have an ORF and is monoallelically expressed in normal kidney, but is biallelically expressed (and its DMR hypomethylated) in Wilms' tumours (Malik et al., 2000). The *WT1-AS* DMR has therefore been suggested to be a primary site of epigenetic disturbance in the pathogenesis of Wilms' tumour.

1.1.14 THE EPIGENETIC MARK

One of the most important features of genomic imprinting is its reprogramming, which occurs upon each passage through the germline in successive generations. How exactly this occurs relates intimately to the question: how are the parental alleles of imprinted genes distinguished or "marked"? Some aspects of this process have been touched upon in the discussion of the molecular basis for PWS and AS. Although the mechanism of imprint establishment is still poorly understood, there is a minimal set of criteria that this process should comply with. These are (Bartolomei and Tilghman, 1997, Ueda et al., 2000):

i) The imprint mark must be established in the gametes, since this is the time when the alleles are in separate compartments and can be differentially modified.

ii) The imprint mark must be stably maintained after fertilisation, in all tissues in which the imprint is recognised.

iii) The modification must be capable of being erased and reset during the production of germ cells, so that the correct sex-specific imprint is transmitted to the progeny.

Considering these criteria, two models have been proposed for the way in which the imprint is set during successive generations. One model proposes that the imprint mark of the same-sex chromosome is simply maintained during gametogenesis, whereas the imprint of the opposite sex is reversed. The second model proposes that existing epigenetic modifications are first erased from both parental chromosomes in both germ lines, and that the imprints are then established in a sex specific fashion at a later stage (Constancia et al., 1998).

Since, as described above, it is well established that methylation plays an important role in genomic imprinting, part at least of the differential methylation present in imprinted genes has been considered as the strongest candidate for the imprinting mark. In mouse germ cells, global demethylation occurs by E12.5-E13.5, when all imprinted sequences are demethylated in both sexes. This is followed by a remethylation wave at E15.5, only certain sequences (CpG sequences) retaining methylation differences between the oocyte and sperm genomes (Razin and Shemer, 1995). The global demethylation occurring at early stages of germ cell development also includes imprinted genes, but it has been demonstrated that *H19* and *Igf2* are not completely demethylated, and that slightly higher levels of methylation are present in male compared to female primordial germ cells (PGCs). However, it has also been shown that *Igf2* and *H19* (and also *Igf2r* and *Snrpn*) are biallelically expressed in both germlines, beginning by E11.5. This suggests that their imprints are indeed largely erased. Interestingly, one region that escapes the global demethylation wave seems to be the 5′ end of *Xist*, as it has been shown that it remains methylated in both sexes at this stage (Villar et al., 1995, Szabo and Mann 1995, Razin and Shemer, 1995, Ben-Porath and Cedar, 2000, Ueda et al., 2000).

After imprint erasure, the second step in this process is the re-establishment of imprinting. This was studied by Chaillet et al. (1991) using a transgene locus. It was found that the differential methylation patterns of this imprinted transgene were erased in the PGCs and that distinctive patterns emerged afterwards, during germ cell maturation.

In the female mouse, the oocytes in dictyote stage arrest (from E13.5) are apparently not methylated until after birth, when methylation occurs during oocyte growth. This process coincides with the presence of high levels of *Dnmt1* in the nucleus of the growing oocyte (Constancia et al., 1998).

After fertilization, the remodelling of methylation occurs at sites flanking the core DMR from very early stages, ultimately producing the allelic methylation state of somatic tissues. In one sense, at their CpG islands the two alleles must withstand opposing modifying processes, in that methylation on one allele must be protected from the general demethylation occurring over the rest of the genome and also the hypomethylated state of the other allele must resist the subsequent global remethylation wave. Transcription factors (such as Sp1) seem to participate in the protection of CpG islands from *de novo* methylation. Transcription may thus have a role in keeping CpG islands in their hypomethylated state, presumably by preventing their *de novo* methylation or by inducing demethylation (Razin and Shemer 1995, Constancia et al., 1998).

Finally, the states of demethylation and remethylation must depend at least partly on the activity of the different methyltransferase enzymes. Ramsahoye et al. (2000) have proposed that restoration of DNA methylation after blastocyst implantation depends on the *de novo* activity of *Dnmt3* and its expression in early development indicates that it is the major enzyme responsible for re-establishing the methylation patterns. In early postimplantation development *Dnmt1* supports the methylation carried out by *Dnmt3* by completing the methylation of those CpG sites that have been hemimethylated by *Dnmt3* and by maintaining methylation patterns through replication. When *Dnmt3* is down-regulated in later development, no further *de novo* methylation occurs. The CpG specificity of methylation in the soma reflects the specific maintenance methyltransferase activity of *Dnmt1* (Ramsahoye et al., 2000).

1.1.15 EVOLUTION OF GENOMIC IMPRINTING

One of the most intriguing questions regarding genomic imprinting is "why is this mechanism necessary at all?" Imprinting must confer some evolutionary advantage, to counteract the potentially deleterious consequences of the fact that by silencing one allele of certain genes, mammals have discarded the advantages of diploidy in exchange for a state of haplosufficiency (Preece and Moore, 2000).

The three most current hypotheses concerning the evolution of genomic imprinting suggest that this process may have: i) evolved as an extension of the host defence mechanism of prokaryotes, or ii) evolved as a mechanism to protect female mammals from malignant trophoblast disease or finally iii) originated to control fetal growth and also natural selection (Hall, 1990, Preece and Moore, 2000).

i) The host defence hypothesis was suggested by Barlow (1993), based on the knowledge that in bacterial restriction systems, methylases modify the specific recognition sites on the bacterial DNA, thus protecting the host genome from cleavage by restriction endonucleases. Foreign (eg bacteriophage) DNA is not protected and can therefore be digested. By analogy, in mammals the DNA methyltransferases may have evolved to prevent expression of invading viral genomes. If so, genomic imprinting may be a result of the evolution of DNA methyltransferases from an agent for host defence towards a mechanism that regulates gene expression. Nonetheless, this theory does not offer an immediate explanation for why mammals in particular use this machinery to silence a subset of their own genes (Bartolomei and Tilghman, 1997).

The other two hypotheses both concern two circumstances pertaining to mammalian reproduction. One is that mothers have to tolerate the invasive characteristics of a tissue which is half "foreign" and the second is that they must be able to restrict the growth potential of their offspring in order not to sacrifice maternal well-being (and/or the well-being of other subsequent litters) (Hall, 1990).

ii) Varmuza and Mann (1994) proposed that genomic imprinting evolved as a mechanism to protect female mammals from malignant trophoblast disease. These authors considered that as in no other group of animals does the embryo invade its parent tissue so intimately, female mammals gain protection from malignant trophoblast disease by the inactivation in their oocytes of those genes necessary for trophoblast development. (As mentioned elsewhere, paternal genes are known to be required for trophoblast development.) However, this theory does not explain why there is silencing of some paternal genes (Bartolomei and Tilghman, 1997).

iii) A third theory (and one that has gathered more experimental support) was proposed by Moore and Haig (1991). It suggests that imprinting evolved as result of the conflicting "interests" of maternal and paternal genes regarding their offspring. Each parent has different interest in relation to the growth of the offspring (particularly in a polygamous species). The male's interest is in getting the female to invest as much nutritional resources as possible in his offspring. Paternally derived genes would thus foster large offspring, while maternally derived genes would evolve in such a way as to moderate growth, in order to safeguard the mother and her future reproductive potential. As it seems that monogamy is comparatively rare in animals, it has been suggested that imprinting emerged as an evolutionary consequence of patterns of reproduction in placental mammals (Haig and Graham, 1991, Barlow, 1995, Bartolomei and Tilghman, 1997, Pennisi, 1998, Pagel, 1999).

In the context of the relationship between mother and offspring in mammalian development, it may be noted that disruption of the maternally imprinted gene *Peg1/Mest1* resulted in general retardation of embryonic growth, but in the particular case of adult females, was also associated with poor maternal care. This suggests that imprinted genes may extend their influence on growth regulation into the adult stage, thereby affecting the care of the next generation (Lefebvre et a., 1998, Nicholls, 2000).

Insights into the validity of the parental conflict theory for evolution of genomic imprinting may also be had from comparative studies, of for example the *cis*-

regulatory mechanisms in humans and other species, e.g. marsupials and monotremes. Monotremes (that include the platypus and spiny anteater) have been considered as the earliest mammalian lineage, preceding the later divergence of marsupials and eutherian mammals. There are two theories regarding the evolution of these groups, under one of which monotremes and marsupials belong to a shared clade, while the other suggests an early split of monotremes from the marsupial-eutherian clade (John and Surani, 2000).

The *M6P/IGF2R* gene is maternally expressed in the South American opossum (a marsupial), as it is in eutherian mammals, although it lacks the intron 2 DMR postulated to be involved in imprint control. *M6P/IGF2R* genes are not imprinted in monotremes, in either platypus or echidna (Killian et al., 2000). It is suggested that these results indicate that invasive placentation and gestational fetal growth are not required for imprinted genes to evolve. Unless there was convergent evolution of M*6P/IGF2R*, these results also support the view that the lack of genomic imprinting in monotremes reflects the fact that these species represent the earliest mammalian forms, and that genomic imprinting originated later in the marsupial and eutherian linages (Killian et al., 2000).

However if the view of a shared monotreme-marsupial clade prevails, then it must be considered that genomic imprinting in eutherians and marsupials may have evolved independently, unless there was a subsequent loss of this form of gene regulation in monotremes (John and Surani, 2000, Killian et al., 2000).

Considering once again the parental conflict theory, it has been suggested that imprinting would not evolve in monogamous species, because the parents will have a common interest in all their offspring. However, in the monogamous mice *P. polionotus* and other species, *H19, Igf2r* and *Igf2* are imprinted. This situation could be explicable if even limited partner exchange were sufficient to induce conflict, but it should also be considered that the evolutionary distance between the different species investigated is not great enough for a loss of imprinting to have occurred (Pennisi, 1998, Tilghman, 1999).

It has been demonstrated that genomic imprinting also occurs in plants, as in the case of the gene *MEDEA* (*MEA*) in *Arabidopsis*. *MEA* was the first gene relevant to seed development demonstrated to be imprinted. It is maternally expressed in the embryo sac; however given that *Arabidopsis* self-pollinates, all the offspring of a given plant have the same father, and so the parental conflict should be relaxed. By way of explanation, it has been suggested that imprinting in *Arabidopsis* is an evolutionary relic inherited from outbred ancestors, and is in the process of breaking down, with slight differences among races (Mora-Garcia and Goodrich, 2000). It is also of interest that imprinting is not observed in birds; in these species once the egg is laid the resources are fixed, and can consequently no longer be altered by the action of paternal genes (Devriendt, 2000).

Although there is evidence that supports all three hypotheses regarding the evolution of genomic imprinting, none of them is an entirely satisfactory explanation for this phenomenon. It may be that all of them reflect different parts of the role and the importance that this mechanism has regarding gene regulation and evolution.

The experimental work described in the remainder of this thesis relates to efforts to investigate the imprinting of an imprinted region not discussed above, the *GNAS1* region of human chromosome 20, and to determine the extent of this imprinted domain. Further background material that led up to this work is presented at the start of Chapter 3.

CHAPTER 2

MATERIAL AND METHODS

CHAPTER 2. MATERIAL AND METHODS

2.1 MATERIALS

2.1.1 Chemicals and reagents

Chemicals for general use were supplied either by Sigma (Sigma-Aldrich Company Ltd.) or BDH (Merck, Ltd.) and were of molecular biology grade. Bacterial media were supplied by Difco U.K. Ltd.

2.1.2 Radiochemicals

The radiolabelled nucleotides [α^{32}P] dCTP at a specific activity of 3000 Ci/mmol and [γ^{32}P] ATP at a specific activity of 1415 Ci/mmol were both supplied by Amersham Life Science. The four Redivue^{TM} ^{33}P-labelled dideoxynucleotide triphosphate terminators were also supplied by Amersham Life Science.

2.1.3 Vectors and markers

The 1 kb ladder DNA marker and other DNA molecular markers used were from Gibco BRL Life Technologies. The vectors used for the construction of subclone libraries were obtained as follows: the pUC18 vector was obtained from Stratagene, the pCRII vector from Invitrogen and the TVEC vector from Promega. The Epicurian Coli ® JM109 competent cells were purchased from Stratagene.

2.1.4 ENZYMES

Restriction enzymes and T4 DNA ligase were supplied either by Gibco or by New England Biolabs. *Taq* DNA polymerase was supplied at a concentration of 5U/µl by Gibco and Promega. ThermoSequenase^{TM} DNA polymerase was supplied in the ThermoSequenase radiolabeled terminator cycle sequencing kit and in the

ThermoSequenase™ dye terminator cycle sequencing pre-mix kit, both from USB/Amersham Life Science.

2.1.5 SOLUTIONS AND BUFFERS

Solutions and buffers were prepared using distilled and deionised water and were stored at room temperature (between 15°C and 25°C). Sterilisation of the solutions and buffers was performed by autoclaving at 121°C for 15 minutes. The components of general solutions and buffers are described in appendices 2A and 2B.

2.1.6 ELECTROPHORESIS AND DNA TRANSFER MATERIALS

Agarose gels were prepared using standard grade agarose from Sigma Chemical Company Ltd. The 19:1 acrylamide:bisacrylamide mix used for preparing sequencing gels was supplied as a 40% solution by BDH Chemicals. Hybond-N^+ membranes were supplied by Amersham Life Science.

2.1.7. HUMAN GENOMIC DNA

The normal human genomic DNA used in various places during these studies was obtained from 5 non-related female subjects. This panel of normal DNAs was used, among other things, for the identification of expressed polymorphisms in the 3′ utr regions of the *CTSZ* and *CGI-107* genes (Chapters 5 and 6).

2.1.7.1 Parthenogenetic DNA

Human genomic DNA of parthenogenetic origin was obtained from a lymphoblastoid cell line established from the parthenogenetic chimaera F.D (Strain et al., 1995). The leucocytes of this patient had been shown to be 100 % parthenogenetic in origin.

This parthenogenetic DNA was used for various purposes, based on the central idea that the differentially methylated regions (DMRs) associated with imprinted genes

could be identified or highlighted by comparing the methylation patterns of normal and parthenogenetic DNA. Analysis by Southern blotting for known imprinted loci (*SNRPN*, on chromosome 15 and *IGF2R* on chromosome 6) had confirmed that F.D. lymphoblastoid cell DNA lacked paternal-specific methylation patterns, suggesting that screening for methylation differences between F.D. and normal DNA might be used to identify unknown imprinted human genes.

Such a screen for regions displaying differential allelic methylation (DMRs) was carried out in collaboration with the laboratory of Dr. Y. Hayashizaki (Tsukuba, Japan). This screen used restriction landmark genomic scanning for methylation differences (RLGS-M), a technique developed in the Hayashizaki laboratory (Hatada et al., 1991, Hayashizaki et al., 1994, Kamiya et al., 2000). Using the RLGS-M technique, two spots named NV149 and A20 were identified as candidate imprinted loci. These were excised from the gel, and cloned and sequenced. NV149 proved to map upstream of the *ZAC* gene and allowed the demonstration that *ZAC* is a paternally expressed imprinted gene and most likely the gene that is dysregulated in transient neonatal diabetes mellitus (Kamiya et al., 2000). The other spot, A20, proved to contain a new exon of the *GNAS1* gene (Chapter 3).

2.1.7.2 Collection of paired fetal and maternal DNAs

In order to analyse allelic expression of the genes described in this thesis, a collection of paired first-trimester fetal and maternal DNAs was first used, to identify fetuses heterozygous for an expressed sequence polymorphism. Thereafter, reverse transcription of informative fetal RNAs from several tissues was performed. This collection of samples has been described before (Campbell et al., 1994) and includes fetal tissues between 6 to 12 weeks gestation, together with a maternal blood sample.

Maternal and fetal DNAs were extracted from leucocytes and fetal tissues respectively. From fetal tissues, RNA had also been extracted, and using this material, fetal cDNA was prepared. This was done by reverse transcription of total RNA from fetal tissues (see below), and the cDNA mix was then used directly for

PCR amplification. Some of the cDNA samples used in these experiments had been already prepared by Dr. Bruce Hayward for use in previous analyses. 1 µl of cDNA was used for each PCR amplification.

2.2 METHODS

2.2.1 DNA ISOLATION

2.2.1.1 Isolation of plasmid DNA

A protocol modified by R. Treisman from the original method by Birnboim and Doly, 1979, and Ish-Horowicz and Burke, 1981, was used for isolation of plasmid DNA. A single colony was used to inoculate 10 ml of LB-broth, containing the appropriate antibiotic. The culture was incubated overnight at 37°C, using vigorous shaking (225 rpm). The bacteria were pelleted by centrifugation at 4000 rpm for 10 min, and resuspended in Solution I (500 µl) (Appendix 2A). 1 ml of Solution II was added and mixed, followed by incubation at room temperature for 5 to 10 min. Ice cold Solution III (750 µl) was added and the mix was shaken vigorously. The lysate was spun down at 4000 rpm for 10 minutes and the supernatant removed into fresh tubes and phenol/chloroform extracted as described below (section 2.2.1.5).

After the final extraction, the aqueous phase was transferred into fresh tubes and the DNA was precipitated by adding 2 vol of ice-cold absolute ethanol. The tubes were centrifuged at 4000 rpm for 10 min. The supernatant was discarded and the DNA pellet washed with 70% ethanol and allowed to air dry. The pellet was resuspended in 200 µl TE containing DNAse-free RNAse (at 20µg/ml), and stored at -20°C.

The quality of DNA obtained by this method was adequate for sequencing using the ^{33}P cycle sequencing method and for PCR analysis. A better quality of DNA for fluorescent cycle sequencing was obtained when the S.N.A.P.™ kit (a *s*imple *n*ucleic *a*cid *p*rep) kit by Invitrogen was used. The DNA obtained with this kit was used in both ^{33}P and ABI automatic sequencing.

2.2.1.2 Isolation of plasmid DNA with the S.N.A.P.^{TM} kit

The S.N.A.P.^{TM} kit was used mainly to obtain plasmid DNA of good quality for sequencing by the ThermoSequenase^{TM} dye terminator cycle sequencing method. The kit components were resuspension buffer, lysis buffer, precipitation buffer, binding buffer, wash buffer and final wash buffer.

A single colony was used to inoculate 3 ml of LB-broth, containing 20 μg/ml ampicillin, which was incubated with vigorous shaking (225 rpm) at 37°C overnight. The bacteria were pelleted in a microcentrifuge at 14000g for 5 minutes, and resuspended in 150 μl of resuspension buffer by gently pipetting up and down. 150 μl of lysis buffer was added and mixed gently by inverting the tube 5-6 times. After 3 minutes at room temperature, 150 μl of ice-cold precipitation salt was added and the tube was inverted 6-8 times to ensure thorough mixing. The mixture was centrifuged at 14000g for 5 minutes, after which the supernatant was pipetted into a sterile microcentrifuge tube and the pellet was discarded. 600 μl of binding buffer was then added and after mixing by inversion, the solution was pipetted onto the S.N.A.P. ^{TM} miniprep column, which was then placed into a 2 ml collection tube.

The plasmid DNA was bound to the column by centrifuging the solution through it for 30 s. The column was then washed with 500 μl of wash buffer, followed by 900 μl of final wash buffer. The column was then dried by brief centrifugation and finally the plasmid DNA eluted in 60 μl of TE or dH_2O.

Alternatively, when a larger amount of high quality DNA from a specific plasmid was needed, Qiagen Maxi kits were generally used, according to the suppliers' instructions.

2.2.1.3 DNA isolation from P1 artificial chromosome (PAC) clones

PAC DNA was isolated by the rapid alkaline lysis miniprep method, a modification of a standard Qiagen Tip method that uses no organic extractions or columns (as suggested by the UK HGMP Resource Centre and as adapted for Pieter de Jong's laboratory). PAC clones of interest were received from the HGMP Resource Centre on LB-agar slopes containing kanamycin. To obtain single colonies, they were restreaked onto the same agar type, and incubated at 37°C overnight.

A single isolated bacterial colony was used to inoculate 2 ml LB-broth supplemented with 25µg/ml kanamycin in a 12-15 ml snap-cap polypropylene tube. This culture was incubated at 37°C overnight (up to 16 h) with vigorous shaking (225 rpm). The cells were pelleted by centrifugation at 3000 rpm for 10 min. The supernatant was discarded and the pellet was resuspended in 0.3 ml of P1 solution (see Appendix 2A). 0.3 ml of P2 solution was then added and the tube was gently shaken. After 5 min at room temperature, the appearance of the suspension changed from very turbid to almost translucent. 0.3 ml of P3 solution was slowly added to each tube and gently shaken during addition. A thick white precipitate of protein and *E. coli* DNA was formed.

After at least 5 min on ice, the tube was centrifuged at 10000 rpm for 10 min at 4°C, and returned to ice. The supernatant was transferred to a 1.5 ml Eppendorf tube that contained 0.8 ml ice-cold isopropanol. The sample was mixed by inverting the tube a few times and placed again on ice for at least 5 min before centrifuging in a cold microcentrifuge for 15 min. The supernatant was removed and 0.5 ml of 70% ethanol was added; the tube was inverted several times to wash the DNA pellets. This procedure was performed twice, after which the supernatant was removed and the pellet was allowed to dry at room temperature (15-25°C) before being resuspended in 40µl TE.

2.2.1.4 DNA isolation from P1 artificial chromosome (PAC) clones; modified method

Although the quality of the PAC DNA obtained with the above method was adequate for sequencing using ^{33}P terminators, it was found that the quality and quantity of PAC DNA obtained could be improved by using a modified version of the alkaline lysis method.

A single colony was used to inoculate 5ml of LB containing kanamycin (25µg/ml), which was incubated at 37°C for 4-7 hrs with vigorous shaking (225 rpm). This culture was used to inoculate 250ml LB for overnight culture at 37°C with vigorous shaking (225 rpm). The bacteria were pelleted by centrifugation at 6000 rpm for 20 min and resuspended in 5 ml of solution P1. 5 ml of solution P2 was then added and mixed (this was done in Corex™ tubes). The mix was left at room temperature for 10 min, 5 ml of solution P3 was added, and the tube was placed on ice for 10 min.

The cell debris was removed by centrifugation at 12000 rpm for 20 min at 4°C. The supernatant was filtered through cheesecloth into clean Corex™ tubes and ice cold isopropanol (0.6 vol.) was added. After mixing well, the mixture was centrifuged at room temperature at 12000 rpm for 20 min. The supernatant was poured off and the pellet was washed with 70% ethanol and allowed to air dry, before being resuspended in1.5 ml TE (pH8). 1.5 ml ice cold 5M LiCl was added and the mixture was centrifuged at 12000 rpm for 15 min at 4°C. The supernatant was transferred to fresh Corex™ tubes and an equal volume of isopropanol was added. After mixing the mixture was centrifuged at 12000 rpm for 20 min and the supernatant was discarded. The pellet was washed with 70% ethanol.

After air drying the pellet was resuspended in 250µl TE (pH8) containing Rnase (20µg/ml) and left at room temperature for 30 min. To this solution was added 250µl 1.6M NaCl containing 13% (w/v) polyethylene glycol. This was mixed well and the PAC DNA was recovered by centrifugation at 12000 rpm for 15 min at 4°C. The

supernatant was removed by aspiration and the pellet was resuspended in 200µl of TE (pH8.0).

2.2.1.5 Measurement of DNA concentration

The concentration of DNA was determined by measuring the absorbance of the solution at 260 nm (A_{260}), using an appropriate water or TE blank as reference. For most purposes, 10µl of sample was diluted in 1 ml of water or TE before measuring. The original DNA concentration was then calculated according to the formula

[DNA] = Dilution factor x A_{260} x 50 µg/ml

2.2.1.6 Phenol/chloroform extraction of DNA

An equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) was added to the DNA solution of interest. After mixing well, the mixture was microcentrifuged for 5 min. The aqueous phase was removed and the procedure was repeated. The second aqueous phase was then transferred to a clean tube and an equal volume of chloroform was added. After microcentrifuging again for 5 min, the aqueous phase was transferred to a new tube and ethanol precipitated.

2.2.1.7 Ethanol precipitation of DNA

The volume of a given DNA sample was measured and the salt concentration adjusted by adding 1/10 volume of sodium acetate (3M; pH 5.2). After mixing well, 2 volumes of ice-cold ethanol were added. When small amounts (generally < 1 µg) of DNA were precipitated, 10 µg of tRNA was added to increase recovery from the dilute DNA solutions. The DNA was placed on ice for 30 min, and then pelleted by centrifugation at 12000g for 15 min. The supernatant was discarded and the pellet was washed twice with 70% ethanol, spinning briefly. The pellet was then allowed to air dry (vacuum drying was also sometimes used) and finally resuspended in the appropriate volume of TE or water.

2.2.1.8 First strand cDNA synthesis

To synthesize the cDNA used in the RT-PCR experiments described in Chapters 3, 5 and 6, the preferred reverse transcriptase was SuperScript II (Gibco-BRL). The RT reaction was set up as follows: 5 µl of 3 µM primer (oligo(dT)RACE), 3 µl of RNA and 10 µl of dH_2O, were combined in an Eppendorf tube on ice, covered with mineral oil, incubated at 85°C for 3 minutes to denature RNA secondary structure and returned to ice for 2 minutes. 6 µl of 5 X RT II Buffer, 3 µl 0.1 M DTT, 1.5 µl 20 mM each dNTP, 0.5 µl RNA guard™ and 1 µl of SuperScript RT II were added. The RNase inhibitor RNA guard ™ was derived from human placenta and was purchased from Amersham Pharmacia at a concentration of 29700 units/ml. The RT reaction was then incubated at 42°C for 60 minutes, 50°C for 15 minutes and 95°C for 15 minutes, and stored at -20°C. For each experiment, 2 µl of this RT reaction were used in a 50 µl final volume PCR, as described in the following section.

2.2.2 THE POLYMERASE CHAIN REACTION

PCR reactions were carried out in a Hybaid Omnigene™ Thermal Cycler or in a Perkin Elmer Thermal Cycler TC1. The reactions were carried out in either 0.5 ml tubes or 96 well microtitre plates. The Gilson pipettes, tips, tubes and reagents used were sterile and kept separately from those used in other experiments. To avoid the possibility of extraneous DNA contamination in the PCRs, all the reactions were set up in a separate room, specially designated for this activity.

The PCRs were carried out in 1x PCR buffer (see Appendix 2A for different recipes used), with 200µM deoxynucleoside triphosphate (dNTPs), 10-20 pmol primers, 1U *Taq* polymerase, and template DNA (0.1 µg genomic DNA, 0.1-1 ng plasmid DNA) in a final volume of 50µl. In order to prevent evaporation and subsequent alteration of the buffering conditions, the PCRs were overlaid with mineral oil (Sigma). Depending on the characteristics of the template, an initial denaturation step at 94°C for 5 min was used followed by the addition of *Taq* polymerase. The PCR cycling

conditions used were in general, 94°C for 1min, the appropriate primer annealing temperature for 1 min and primer extension at 72°C for 1 min (if the product was < 500 bp) or 3 min (if the product was > 500 bp). The overall number of cycles for the reactions was dependent on the type of DNA template; in general 25-30 cycles were used when working with plasmid DNA and 35-40 cycles when working with human genomic DNA and PAC DNA.

Some of the subclone libraries containing DNA fragments derived from the PAC DNAs under study were analysed by colony PCR. The subclone glycerol stocks were kept in microtitre plates and 1 µl of the stock was used directly in the PCR reaction. Alternatively, when the clone of interest was streaked on LB agar plates, a small amount of the colony was taken using an autoclaved pipette tip and transferred to the wall of the PCR tube (in sufficient amount to be visualised) below the liquid surface.

2.2.2.1 Oligonucleotides

The oligonucleotide PCR primers described were generally designed by selecting a sequence of about 20 nucleotides with the aid of one of two computer programs, either "Oligo 4.0" (used formost of the primers) or "Primer 3". The Oligo 4.0 primer analysis software by Wojciech Rychlik takes into consideration among other parameters that the two primers are of approximately equal predicted T_m and that their 3′ ends are not complementary.

The Primer 3 software was developed at the Whitehead Institute in 1998 by Steve Rozen and Helen Skaletsky. This program predicts the best pair of primers to amplify a given target within the desired sequence, taking into account parameters such as primer T_m, pair complementarity and primer size.

Custom-synthesized oligonucleotides were purchased from Sigma-Genosys Ltd., Oswel DNA Services, or MWG-Biotech GmBH. They were supplied as lyophilised products from a synthesis scale of 0.03 µmol. The sequences of oligonucleotides used in the work described in this thesis are listed in the corresponding chapters.

2.2.3 RESTRICTION ENZYME DIGESTION OF DNA

Restriction endonuclease digestions were performed under the buffer and temperature conditions recommended by the suppliers for each enzyme. In general terms, depending on the amount of DNA used, reactions were carried out in final volumes of 20, 50 or 100 µl (made up with sterile water) and incubated for 5-16 hr. Typically 1 µg of plasmid and PAC DNA was digested, while in the case of genomic DNA 10 µg was used. If required for further manipulation of the digested DNA, the sample was then phenol/chloroform extracted and ethanol precipitated.

2.2.4 AGAROSE GEL ELECTROPHORESIS

DNA was usually size fractionated and visualised by agarose gel electrophoresis. Gels were prepared at concentrations ranging from 0.8 to 3% agarose (weight/volume) in the appropriate volume of 0.5 x TBE buffer or 1 x TAE (see Appendix 2A). In general terms, higher percentage gels were used to visualise low molecular weight DNA, and low percentage gels were used to visualise high molecular weight DNA. Ethidium bromide was added to a final concentration of 0.5 µg/ml to molten agarose before each gel was poured. Before electrophoresis, DNA samples were combined with one-tenth volume loading buffer (see Appendix 2A). The samples were then electrophoresed alongside an appropriate size marker at 50-150 volts until the dye marker had run a sufficient distance. The DNA was visualised by viewing under ultraviolet (UV) transillumination at 260-270 nm and photographed using an electronic imager (Fujifilm thermal imaging system FTI-500, ImageMaster®VDS, Pharmacia Biotech). Where fragments were to be purified from the gel for further manipulation, a 360 nm hand-held UV lamp was used to visualise the DNA, in order to minimise photolytic damage to the DNA.

2.2.5 ALKALI BLOTTING

The genomic DNA (~10 μg) or PAC DNA (~1 μg) of interest was digested with the desired restriction enzyme(s) in a final volume of 50 μl. The digests were separated by electrophoresis on a 0.8% TAE agarose gel and transferred to a positively charged nylon membrane (Hybond N+). Prior to transfer the gels were depurinated in 0.25 M HCl for 15 minutes and then blotted using 0.4 M NaOH as transfer buffer. The transfer apparatus consisted of a transfer buffer reservoir, a blotting platform with 3MM filter paper wicks, and paper towels above the gel. Gels were placed on the blotting platform on top of three layers of 3MM filter paper and surrounded with cling film. The gel was then sequentially overlaid with the Hybond N+ membrane, three further layers of 3MM paper and a 5-10 cm stack of absorbent paper towels. Pressure was applied using a rigid glass or perspex plate with a weight of ~ 200 g. The transfer was allowed to proceed for at least 16 h, after which the membrane was rinsed in 2 x SSC and air-dried. The DNA was fixed to the membrane by baking at 80°C and/or by cross-linking (using a Stratalinker® UV crosslinker).

2.2.6 PREPARATION OF SUBCLONE LIBRARIES

Subclone libraries were prepared from several PAC DNA clones. The plasmid vectors used were pUC18, pCRII and TVEC. Typically, 1 μg of PAC DNA was digested with 50 U of enzyme in a final volume of 100 μl for 3 hours at 37°C. The digested DNA was phenol/chloroform extracted, and ethanol precipitated with 10 μg of carrier tRNA. The pellet was resuspended in 10 μl of dH_2O at a nominal concentration of ~100 ng/μl.

The vectors used for the preparation of the subclone libraries were digested with the corresponding restriction enzyme and the phosphate groups were removed by treatment with shrimp alkaline phosphatase to prevent religation of the vector to itself.

200 ng of insert DNA were ligated to 40 ng of vector overnight at room temperature, using 1U of T4 DNA ligase in a 20 µl final volume containing T4 DNA ligase buffer (included in the T4 ligase kit). The ligation reaction was then ethanol precipitated and the final pellet was resuspended in 4 µl of dH_2O.

2.2.7 PREPARATION OF ELECTROCOMPETENT *E. coli*

The host cells used were *E. coli* strain JM109 purchased from Stratagene. A culture of JM109 was grown for 16 h at 37°C in 10 mls of LB-broth. This culture was used to inoculate 1 l of sterile LB-broth, which was incubated at 37°C with vigorous shaking, until the apparent optical density (A_{600}) reached 0.5-1. The culture was chilled on ice for 30 min and then pelleted by centrifugation at 4000 g for 15 min. The supernatant was removed and the bacterial pellet was resuspended in 1 l of ice-cold sterile dH_2O. The cells were pelleted and washed again in 20 ml of ice-cold 10% glycerol. The cells were then frozen in 45 µl aliquots and stored at -70°C.

2.2.8 TRANSFORMATION OF COMPETENT *E. coli*

For transformation of *E. coli* by electroporation, 1 µl of the precipitated, resuspended ligation reaction was used. This was added to the 45 µl bacterial aliquot and left on ice for at least 5 min. The cells were placed in a chilled 0.2 ml electroporation cuvette and then pulsed using a GenePulser™ apparatus (BioRad), set at 25 µF, 2.50 kV, with the external pulse controller at 200Ω. . The cuvette was removed from the apparatus and 1 ml of LB-broth (without ampicillin) was immediately mixed with the cells, which were then cultured at 37°C for 1 h to allow pre-expression of the antibiotic resistance. Thereafter, 200 µl of this culture were plated onto LB-agar plates (see Appendix 2B) containing appropriate antibiotic selection and cultured at 37°C for 16 h.

The subclone colonies obtained from these transformations were picked into 200 µl of LB-broth and ampicillin in 96 well plates. These were incubated overnight and then glycerol added to 10% final concentration for storage at -70°C until required.

2.2.9 PREPARATION OF THE HYBRIDISATION MEMBRANES USED FOR SUBCLONE LIBRARY SCREENING

To allow several rounds of future screening by hybridisation, the subclone libraries were replicated onto hybridisation membranes. Cultures were transferred to nylon membranes laid on LB-ampicillin plates using a 96-prong "hedgehog" plating device. They were allowed to grow for 12-16 hours before lysis and fixing of the DNA. These cultures were performed in 22x22 cm plastic plates with lids, which were prepared by washing with detergent, rinsing twice with dH_2O, drying, washing with 70% ethanol, drying again and sterilisation by exposure to UV light for 20 min. 250 ml of molten LB-agar-ampicillin was poured into each plastic plate and allowed to set, and the plates finally exposed again to UV light as above before use.

Each prong of the stainless steel "hedgehog" plating device corresponds to one well of the 96-well plates used to store glycerol stocks, and also has a notch at its end that by capillary action retains a few µl of the thawed culture. Care was taken to avoid contamination of the glycerol stocks when more than one plate was being transferred; prior to dipping into the wells, the spikes were rinsed three times in dH_2O and then sterilised twice with 70% ethanol and a Bunsen burner. After sampling the glycerol stock plate, the cultures were immediately stamped out onto a 20 x 20 cm hybridisation membrane twice, side by side, to create a master and a duplicate grid.

After the culture period, the membranes were first submerged in 10% SDS for 5 min, then in denaturing solution (see Appendix 2A) for 5 min and then in neutralising solution for 15 min. They were allowed to dry on 3MM filter paper, and the DNA was then fixed to the membrane by baking at 80°C for two hours. Finally, prior to

hybridisation the membranes were washed in 2 x SSC to remove any remaining bacterial debris.

2.1.10 HYBRIDISATION SCREENING OF THE SUBCLONE LIBRARIES

Probes for hybridisation were generated by a modification of the random priming method (Feinberg and Vogelstein, 1984) using the T7 QuickPrime^{TM} kit (Amersham Pharmacia Biotech) with [α^{32}P] dCTP (3000Ci/mmol) as the radionucleotide. The DNA fragment for labelling (25-50 ng) was isolated on a low-melting-point agarose gel. The band was excised from the gel with a minimum of excess agarose and placed in a pre-weighed microcentrifuge tube that was then weighed again. 3 ml of distilled water per gram of gel slice was added and the tube was heated at 65°C for 2 minutes to melt the agarose. The sample was mixed to homogeneity and 25 µl aliquots were prepared. Prior to labelling, each aliquot was heated in a water bath at 95°C for 7 minutes.To a clean microcentrifuge tube was added the denatured DNA, reagent mix (a buffered aqueous solution containing dATP, dGTP, dTTP and random 9mer oligodeoxyribonucleotides), 3 µl of [α^{32}P] dCTP (30µCi) and 4 U of T7 DNA polymerase. The components were mixed gently and then centrifuged briefly and incubated at 37°C for 15 minutes. The labelled DNA was denatured by heating at 95°C for 2 minutes, then cooled immediately on ice and added to the hybridisation bottle.

For hybridisation, the membranes were rolled between nylon nets, and applied to the inner surface of cylindrical hybridisation bottles. 20 ml of prehybridisation mix (see Appendix 2A) was added to the hybridisation bottle, which was rotated at 65°C for 1 h. Afterwards, the prehybridisation mix was replaced by 10 ml of prewarmed hybridisation mix containing labelled probe as above and the hybridisation was allowed to continue overnight.

For the 5′ end labelling of oligonucleotides with γ^{32}P-ATP, the labelling reactions consisted of 30 ng of primer per sample, 1 µl of 10 x PNK buffer, 3 µl of γ^{32}P-ATP, 5U of T4 polynucleotide kinase (PNK) and dH_2O to 10 µl. The reaction was

incubated at 37°C for 40 min and then added directly to the hybridisation bottle. The hybridisation in this case was allowed to proceed overnight at 48°C. The hybridisation mix used was the Quick oligo hybridisation mix (see Appendix 2A).

2.1.11 POST-HYBRIDISATION WASHING AND RADIOACTIVE SIGNAL DETECTION

After hybridisation, non-specifically bound random-primed DNA probe was removed by washing the membranes twice in 2 x SSC/0.1% SDS for 20 min at 65°C and once in 1 x SSC/1% SDS for 20 min at 65°C. Afterwards, the membranes were sealed in a plastic bag or cling film and exposed to X-ray film in light-proof cassettes with intensifying screens for between 16 h (overnight) and several days. When oligonucleotide probes and the Quick oligo hybridisation mix were used, the membranes were washed only at a stringency of 4 x SSC/0.1% SDS at 48°C for 4x 5 min. If needed for subsequent hybridisation experiments, the membranes were washed for longer periods of time as before or at greater stringency, and then exposed to autoradiographic film as before to ensure complete removal of probe.

2.1.12 DNA SEQUENCING

DNA sequencing was carried out using one of two commercially available protocols: the ThermoSequenase™ radiolabelled terminator cycle sequencing kit or the ThermoSequenase™ dye terminator cycle sequencing kit (both Amersham USB). ThermoSequenase™ is a *Taq* polymerase derivative that has been modified by site-directed mutagenesis to increase its substrate affinity for dideoxynucleoside triphosphates, thus greatly improving its characteristics for sequencing (Tabor and Richardson, 1995).

Plasmid DNA was used directly in cycle sequencing without further manipulation after extraction. PCR products, however, were first treated with exonuclease I and shrimp alkaline phosphatase (SAP). The exonuclease I was used to remove residual single-stranded primers. The shrimp alkaline phosphatase removes the remaining

dNTPs from the PCR mixture, which would otherwise interfere with the sequencing reaction. Typically, reactions were performed using 5 µl of PCR product, 1µl (10U) of exonuclease I and 1µl (2U) of SAP, for 15 min at 37°C, after which the enzymes were heat inactivated at 80°C for 15 min.

2.1.13 Manual (radioactive) sequencing

The present methods for DNA sequencing are modifications of the original Sanger end-terminator sequencing techniques (Sanger, 1977). They are based on the principle that a 2′,3′-dideoxynucleotide (supplied as either ddGTP, ddATP, ddTTP or ddCTP) can be incorporated into the end of a growing DNA chain by forming a phosphodiester bond between its 5′ carbon atom and the 3′ carbon of the previously incorporated nucleotide. However, its 3′ carbon atom cannot participate in further phosphodiester bonding due to the lack of a –OH group, resulting in a base-specific termination of chain synthesis (hence the name "chain termination").

The termination mixes were prepared for each [α^{33}P] labelled ddNTP (A, C, G, T) according to the supplier's specifications, each containing 0.25µl of each radiolabelled ddNTP. A DNA-primer premix was then prepared, consisting of 1 µl of the nucleotide master mix (containing either dGTP or, when required to eliminate compression artefacts occurring in G/C-rich sequences, dITP), 3.2 pmol primer, 1µl of reaction buffer (see Appendix 2A), 25-250 fmol template DNA, 1µl of ThermoSequenase polymerase, and water to a final volume of 10 µl.

This premix was distributed between the four termination mixes. 40 to 45 cycles of 95°C for 30 s, 55°C 30 s, 72°C 1 min were then performed, after which 4 µl of formamide stop solution was added to each reaction. The samples were denatured for 5 min at 80°C and 2.5 µl of each was loaded immediately onto a glycerol-tolerant 6% polyacrylamide sequencing gel (see Appendix 2A). The gels (0.4 mm x 38cm x 50cm) were prepared in Biorad sequencing gel rigs. The gel solution came from a 950 ml stock to be made up to the equivalent of 1 litre with 20X glycerol tolerant buffer; the other components were 8M urea, 6% acrylamide/N, N'-methylene

bisacrylamide (19:1; Acrylogel 5, BDH-Merck) and dH_2O. To 95 ml of this stock were added 5 ml of 20X glycerol tolerant buffer and the gel was polymerised by adding 150 µl of 25% ammonium persulphate and 100 µl of TEMED immediately before pouring. Polymerisation was allowed to proceed for a minimum of 1 hour, after which electrophoresis was performed using 1X glycerol tolerant buffer at 90 W for 2-6 hours, depending on length of read required. The sequencing reactions were always loaded in the order A, C, G, T.

After electrophoresis the gel was transferred onto 3MM filter paper, covered with cling-film and dried under vacuum at 80°C for ~1 hour. The cling-film was then removed and the gel exposed to autoradiographic film in light-proof cassettes for a minimum of 16 h.

2.1.14 ThermoSequenase™ dye terminator cycle sequencing

This method allows all four termination reactions to be performed in a single tube, since each ddNTP is labelled with a different fluorochrome. Reactions were set up in 0.5 ml PCR tubes, using 4 µl of the terminator ready reaction mix, 100-250 ng double stranded DNA, 6 µM primer, and water to 10 µl. The mix was overlaid with mineral oil and 25 cycles consisting of 96°C for 30 s, 45°C for 15 s and 60°C for 4 min were performed in a Perkin Elmer TC1 themal cycler. The sequencing products were then ethanol precipitated to remove the bulk of unincorporated fluorescent ddNTPs. The pellet was dried at 90°C for 1 min and resuspended in the formamide loading buffer (see Appendix 2A). The samples were denatured at 95°C for 3 to 5 min and kept on ice until loaded onto the sequencing gel.

The sequencing reactions were electrophoresed on an ABI Prism™ 377 DNA sequencer, and the output analysed using the ABI software. This system was used in preference to ^{33}P especially when working with plasmid templates, mainly because it was possible to handle a large number of samples simultaneously. However, for some applications, such as direct sequencing of PAC templates, radiolabelled terminator sequencing was clearly superior in quality.

Various primers, including custom synthesized ones, were used for sequencing as described in the text. For initial characterization of clones, appropriate vector-specific primers were usually used, these being the M13F and M13R primers in the case of the plasmid subclone libraries and the T7 and SP6 primers for the direct sequencing of the ends of PAC clones (see Appendix 2D).

The sequences obtained were generally compared to database sequences using the BLAST program hosted at NCBI (for URL see Chapter 7, bioinformatics) and frequently also the NIX interface for simultaneous multiple sequence analysis programs, hosted at the UK HGMP-RC. BLAST (Altschul et al., 1990) stands for Basic Local Alignment Search Tool and is a set of search programs designed to explore all of the available sequence databases for matches using any combination of either protein or DNA sequences. BLAST uses an algorithm which is able to detect close relationships among sequences which share only isolated regions of similarity.

NIX is a tool used for rapid identification of interesting regions in genomic or transcribed nucleic acid sequences. NIX displays a graphical representation of the ouputs of many programs (including BLAST for example) and allows viewing of these results side by side, making it easy to spot features identified by more than one program. URLs for all the bioinformatics facilities used in this work are listed in the references chapter.

2.1.15 PULSED-FIELD ELECTROPHORESIS

In order to determine the sizes of PAC clones, 1 μg of DNA was digested with at least two rare-cutting enzymes (*Not*I and *Xho*I for example). The digest was analysed on a 1% agarose/TBE gel run without ethidium bromide and using two molecular markers, 1 kb ladder (Gibco BRL) and a MidRange PFG Marker 1 (New England BioLabs). Electrophoresis conditions were 200 volts, pulse time 4 s, temperature 14°C, run duration 16 h. After electrophoresis, the gel was stained with ethidium bromide, and after photography if needed the gels were prepared for blotting as above.

APPENDIX A: Solutions

10% Ammonium persulphate (APS)	10% w/v APS in dH_2O; aliquots were stored at -20°C and defrosted only once
Acrylamide	30% stock (19:1), 28.5 % acrylamide, 1.5% N,N'-methylene bisacrylamide.
Denaturing solution	0.5 M NaOH, 1.5 M NaCl
Denhardt's solution	0.1% Ficoll 400, 0.1% polyvinyl pyrrolidone, 0.1% bovine serum albumin (BSA).
Depurination solution	250 mM HCl
DNA loading dye	0.25% bromophenol blue, 15% Ficoll 400
Ethidium bromide	10 mg/ml in dH_2O
First strand RT II buffer 5 x	250 mM Tris-HCl (pH 8.3), 375 mM KCl, 15 mM $MgCl_2$
Glycerol tolerant buffer 20 x stock (per 10 litres)	Tris-HCl 2160 g, taurine 720 g, EDTA (disodium salt) 40 g, dH_2O to 10 litres
Hybridisation mix	5 x Denhardt's, 5 x SSC, 0.1 % SDS, 0.1% sodium pyrophosphate, 20% dextran sulphate
IPTG	10mM working solution in water
Loading buffer (sequencing)	98% deionised formamide, 10 mM EDTA (pH 8.0), crystal violet 0.01%
Neutralising solution	1.5 M NaCl, 0.5 M Tris HCl, pH 8
6% polyacrylamide sequencing gel	95 ml of acrylamide/urea, 100 µl TEMED 150 µl of ammonium persulphate, 5 mls of 20 x glycerol tolerant buffer
PCR Buffer I (10x stock)	500 mM KCl, 250 mM TAPS, 125 mM KOH, 15 mM $MgCl_2$.
PCR Buffer II (10x stock)	500 mM KCl, 250 mM Tricine, 125 mM KOH, 15 mM $MgCl_2$

PCR Buffer III (10x stock)	100 mM Tris·HCl, 500 mM KCl, 125 mM KOH, 15 mM $MgCl_2$
PCR Buffer IV (10x stock)	500 mM KCl, 250 mM TAPS, 125 mM KOH, 15 mM $MgCl_2$, 1% Triton X-100
PCR Buffer V (10x stock)	500 mM KCl, 250 mM Tricine, 125 mM KOH, 15 mM $MgCl_2$, 1% Triton X-100
PCR Buffer VI (10x stock)	100 mM Tris·HCl, 500 mM KCl, 125 mM KOH, 15 mM $MgCl_2$,1% Triton X-100
PCR 10 x general buffer	15 mM $MgCl_2$, 10 mM Tris·HCl, 500 mM KCl
Phenol:chloroform:isoamyl alcohol	25:24:1 by volume
PNK 10 x buffer	50 mM Tris-HCl (pH 8.0), 10 mM $MgCl_2$' 5 mM DTT
Prehybridisation mix	5 x Denhardt's, 5 x SSC, 0.1% SDS, 0.1% sodium pyrophosphate
Quick oligo hybridisation mix	0.5 g BSA, 0.5 g Ficoll 400, 1 g SDS, 1 g sodium pyrophosphate, 250 ml 20 x SSC, 744 ml dH_2O.
Reaction buffer (manual sequencing)	200 mM Tris-HCl pH 9.5, 65 mM $MgCl_2$
10% SDS	10% w/v sodium dodecyl sulphate. Warm to dissolve, pH to 7.2 with HCl
Solution P1 (filter sterilized, 4°)	15 mM Tris-HCl (pH 8), 10 mM EDTA, 100 µg/ml Rnase A
Solution P2 (filter sterilized, room temp.)	0.2 M NaOH, 1% SDS
Solution P3 (autoclaved, 4°C)	3 M potassium acetate, pH 5.5
Solution I (autoclaved and stored at 4°C)	50 mM glucose, 25 mM Tris-HCl (pH 8.0), 10 mM EDTA (pH 8.0)
Solution II	0.2 M NaOH (freshly diluted from a 10 M stock), 1% SDS

Solution III	5 M potassium acetate 60 ml, glacial acetic acid 11.5 ml, dH_2O 28.5 ml
20 x SSC	3 M NaCl, 0.1 M trisodium citrate, pH to 7
TAE 50 x STOCK (per litre)	242 g Tris base, 57.1 ml glacial acetic acid, 100 ml 0.5 M EDTA (pH 8.0)
TBE 5x STOCK (per litre)	54 g Tris base, 27.5 g boric acid, 20 ml 0.5 M EDTA, pH 8.0
TE 10x stock	10 mM Tris·HCl, 1 mM EDTA, pH 8.0
TEMED	N,N,N',N'-tetramethyl-1,2-diaminoethane, 100% stock
X-GAL	20 mg/ml in dimethylformamide
^{33}P sequencing stop solution	95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol FF

APPENDIX B: Growth Media

Luria-Bertani (LB medium)	10 g/l Bacto-tryptone, 5 g/l Bacto-yeast extract, 5 g/l NaCl,
Luria-Bertani-agar	As before plus 15 g/l Bacto-agar (16 g/l in the case of the 20 x 20 cm plates).

APPENDIX C: Antibiotics

Ampicillin	50 mg/ml stock solution in dH_2O, 20 μg/ml working concentration
Kanamycin	10 mg/ml stock solution in dH_2O, 10 μg/ml working concentration

APPENDIX D: Primers

Name	Sequence
M13F	GTT TTC CCA GTC ACG ACG TTG TA
M13R	AGC GGA TAA CAA TTT CAC ACA GGG
T7	CGC TAA TAC GAC TCA CTA TAG GG
SP6	GTC GAC ATT TAG GTG ACA CTA TA
dT Race	GAG CTC GAG TCG ACA TCG A(T_{17})

CHAPTER 3

GNAS1, AN IMPRINTED GENE IN DISTAL HUMAN CHROMOSOME 20q13

CHAPTER 3. *GNAS1*, AN IMPRINTED GENE IN HUMAN CHROMOSOME 20q13

3.1 INTRODUCTION

The human *GNAS1* gene maps to chromosome 20q13 (Levine et al., 1991). It codes for the α-subunit (Gα) of the heterotrimeric stimulatory G protein, G_s, that activates adenylyl cyclase. Inactivating mutations of *GNAS1* cause the autosomal dominant syndrome pseudohypoparathyroidism type Ia (PHP-Ia) (Levine et al., 1980, 1986).

Pseudohypoparathyroidism (PHP) is a term applied to a heterogeneous group of disorders whose common feature is end-organ resistance to parathyroid hormone (Albright et al., 1942, Patten et al., 1990). In contrast, activating mutations of *GNAS1*, which occur only in mosaic state, cause the McCune-Albright syndrome and have also been found in a number of endocrine neoplasms, notably pituitary somatotroph adenomas (Landis et al., 1989, Cohen and Howell, 1999).

A peculiar pattern of inheritance has been observed in families suffering from PHP-Ia. Within one given kindred, individuals carrying the same inactivating mutation of *GNAS1* may have a different clinical presentation: some individuals display PHP-Ia, while others have no hormone resistance but do have the characteristic skeletal phenotype associated with PHP-Ia (Albright hereditary osteodystrophy). Manifestation of the full PHP-Ia phenotype seems to depend upon maternal transmission of the causative mutation (Davies and Hughes, 1993).

These differences in clinical presentation were considered to be the result of genetic factors that might interact with the *GNAS1* mutations and determine the presentation of the endocrine disturbance. The maternal inheritance pattern led to the suggestion that genomic imprinting might play a role. However, at the start of this study, the evidence for imprinting was conflicting. On the one hand, an RT-PCR based analysis had shown that in a large number of different human fetal tissues *GNAS1* was biallelically expressed (Campbell et al., 1994). On the other hand, *in situ*

hybridization to tissues of mice with uniparental disomy for the chromosomal region including the homologous gene *Gnas* indicated a predominantly paternal origin of glomerular *Gnas* transcripts. This result was also opposite to the pattern of imprinting predicted from PHP-Ia families (Williamson et al., 1996).

This chapter describes more recent studies of human *GNAS1*, in which it was demonstrated that this gene is after all imprinted, as predicted from the anomalous inheritance of PHP-Ia, and that its allele-specific regulation is highly complex.

3.1.1 PARATHYROID HORMONE

In mammals parathyroid hormone (PTH) is the most important regulator of calcium ion homeostasis. The peptide hormone is synthesised as a precursor protein, preproPTH, containing a pre "leader" sequence of 25 amino acids and a pro-sequence of 6 amino acids, which are both cleaved during the synthesis and secretion process to yield the mature form of 84 amino acids (Spiegel and Weinstein, 1995, Mannstadt et al., 1999).

PTH is produced by the parathyroid glands and its synthesis and secretion are largely regulated by the extracellular concentration of calcium. In response to low blood calcium levels, PTH is secreted into the circulation and then acts primarily in kidney and bone. In kidney, PTH directly stimulates the tubular reabsorption of calcium. Elevations in serum phosphate concentration depress serum calcium by reducing bone resorption and by decreasing synthesis of 1,25-$(OH)_2$ vitamin D (Spiegel and Weinstein, 1995, Mannstadt et al., 1999).

A PTH-related peptide known as PTHrP, was initially isolated from human carcinomas and is responsible for the hypercalcemia associated with various malignances. Unlike PTH, however, PTHrP does not circulate in appreciable amounts in normal subjects but is instead, widely expressed in fetal and adult tissues, where it is thought to regulate cell differentiation, cell proliferation, and organogenesis as a paracrine or autocrine soluble factor. It also plays a major role in

chondrocyte maturation and endochondral bone formation (Amling et al., 1997, Mannstadt et al., 1999). It also has been demonstrated that PTHrP is able to increase the expression of Bcl-2, a protein that controls programmed cell death in several cell types (Amling et al., 1997).

Human PTHrP mRNAs encode three different protein isoforms containing 139, 141, and 173 amino acids. The first 13 amino acids of PTHrP and PTH display a high degree of homology (eight are identical). This similarity decreases markedly in the 14-34 region, where only three amino acids are identical, and beyond residue 34 there is no recognizable similarity (Wysolmerski and Broadus, 1994). For both PTH and PTHrP, the 15-34 region functions as the principal PTH1R binding domain. PTHrP binds to the same receptor as PTH, and the biological responses achieved by either ligand through this common receptor are largely indistinguishable, at least regarding mineral ion homeostasis. These actions of PTH (and PTHrP) are mediated through the type 1 receptor or PTH1R; this is a G protein-coupled receptor of the seven transmembrane domain family (Mannstadt et al., 1999).

3.1.2 CELL SIGNALLING UNITS COMPRISING THREE PROTEIN MODULES

A huge number of biochemical functions carried out within eukaryotic cells are regulated by diverse external chemical and physical stimulating factors. These include, for example, a wide range of structurally diverse hormones (catecholamines, gonadotropins, parathyroid hormones, etc.), odorants and light. The control of the various intracellular biochemical functions is achieved by modulating the activities of a small number of signalling units. The individual external effectors do not actually enter the cell, but instead bind to receptors at the cell surface and initiate a "cascade" of intracellular events leading ultimately to a cellular response. During evolution, the repertoire of cell surface receptors has been expanded and adapted to achieve the necessary diversity to respond to so many different external stimuli (Neer, 1995, Sprang, 1997, Berman and Gilman, 1998).

The most important class of signalling units in animal cells are three-protein modules, each consisting of a receptor, an effector and a linking unit. The receptor functions as a signal recognition element; the effectors are signal generators, changes of whose activity cause variation in ionic composition or in second messenger levels (such as cAMP or inositol phosphate) that in turn mediate downstream cellular responses. The activities of these receptor and effector units are linked and coordinated (among other elements) by members of the GTPase superfamily (G proteins) (Berman and Gilman, 1998).

3.1.2.1 G Protein coupled receptors (GPCRs)

Nearly 2000 members of the superfamily of G protein-coupled receptors (GPCRs) have been reported to date. They are classified into over 100 subfamilies according to their sequence homology, ligand structure and receptor function. All GPCRs have a common structural motif containing 7 membrane-spanning regions with their N-terminal end and three interhelical loops to the extracellular side of the membrane, while the C-terminal end and three more loops are cytosolic. N-terminal segments (consisting of 7-595 amino acids), loops (5-230 amino acids), and C-terminal segments (12-359 amino acids) vary in size, an indication of their diverse structure and function (Spiegel, 1996, Ji, et al., 1998, Lodish et al., 2000).

The receptors' intracytoplasmic loops transmit signals from receptor to G protein. Mutational studies have located the essential signal-transmitting information of these loops within short (6-16 residue) amino acid sequences near the junctions with the transmembrane helices (Bourne, 1997). There is evidence that the GPCRs are dynamic proteins that exists in equilibrium between two allosterically different conformations, an inactive (R) or active form (R*) (Bond et al., 1995, Bourne, 1997). Agonists have selective affinity for R*, so that agonist binding shifts the equilibrium in favour of R*. When one such receptor (β_2-adrenoceptor) was highly overexpressed in myocardium of a transgenic mouse, sufficient R* was present to allow maximal signalling response in the absence of agonist (Bond et al., 1995).

3.1.2.2 The superfamily of regulatory GTP hydrolases (G proteins)

The main characteristic of the members of the GTPase protein superfamily is their use of the hydrolysis of GTP to control diverse cellular processes. These proteins can act as "molecular switches": they are turned "on" when bound to GTP and turned "off" when bound to GDP (Schwindinger and Levine, 1997).

Within the GTPase superfamily are the heterotrimeric G proteins and the group of "small" monomeric Ras and Ras-like proteins. These two groups are regulated in very different ways: both classes contain regions that promote the activity of specific effector proteins by direct protein-protein interactions. However, heterotrimeric G proteins are coupled directly to activate receptors, whereas Ras is linked only indirectly via other proteins (Bourne et al., 1990, Koelle, 1997, Schwindinger and Levine, 1997, Lodish et al., 2000).

The group of monomeric proteins includes the Ras, Rho, Rab, Arf, SarI and Ran families. The Ras family regulates gene expression through the MAP kinase cascade; the Rho family regulates reorganization of the actin cytoskeleton. Rab, Arf and SarI families regulate intracellular vesicle trafficking and the Ran family regulates nuclear transport (Sprang, 1997, Sasaki and Takai, 1998).

3.1.2.2.1 Heterotrimeric G proteins structure

All the heterotrimeric G proteins share a common structure consisting of an α-subunit (Gα-subunit) and a tightly coupled βγ-subunit dimer (Gβγ-subunit) (Lodish et al., 2000). The heterotrimeric G proteins were originally classified on the basis of their Gα-subunit characteristics; however the growing number of different Gβ and Gγ-subunits suggests that heterotrimeric G proteins should more properly be defined on the basis of their combination of Gα, Gβ and Gγ-subunits (Rahmatullah et al., 1995).

3.1.2.2.1.1 The Gα subunit of heterotrimeric G proteins

The Gα-subunit protein has several biochemical activities: it binds guanine nucleotides, has intrinsic GTPase activity and couples receptor with effector molecules. Gα-subunits demonstrate the greatest diversity; they are unique to each heterotrimeric G protein and are thought to confer functional specificity to it, allowing discrimination among multiple receptors and effectors. 16 genes encoding Gα-subunit proteins have been reported in mammals; they encode 20 different types of subunits, some of them coding for more than one protein via alternative splicing (Miric et al., 1993, Neer, 1995, Schwindinger and Levine, 1997, Farfel et al., 1999, Lodish et al., 2000).

The Gα-subunits consist of two domains: a GTPase domain that contains the guanine nucleotide-binding pocket as well as sites for binding receptors, effectors and the Gβγ-subunit and an α-helical domain that contributes to the effector-binding site (Masters et al., 1988, Neer, 1995).

3.1.2.2.1.2 Classification of the α subunits of G proteins

Based on the degree of amino acid identity of their α-subunits, G proteins have been divided into 4 subfamilies: G_s (two genes identified), G_i (eight genes), G_q (four genes) and G_{12} (two genes) (Berman and Gilman, 1998). The proteins belonging to each of these groups are shown in Table 3.1.

FAMILY	G-PROTEIN	MW	EFFECTOR	DISTRIBUTION
G_s	$_s\alpha$	45kD	Adenylyl cyclase (+)	Ubiquitous
		52kD	Ca^{2+} channel	
	$_{olf}\alpha$	45kD	Adenylyl cyclase (+)	Olfactory epithelia Sensory neurons
G_i	$_{i}1\alpha$	41kD	Adenylyl cyclase (-) Ca^{2+} and K^+ channels	Widespread brain
	$_{i}2\alpha$	40kD	Adenylyl cyclase (-) Ca^{2+} and K^+ channels	Widespread
	$_{i}3\alpha$	40kD	Adenylyl cyclase (-) Ca^{2+} and K^+ channels	Neuroendocrine
	$_{o}\alpha$	39kD	Ca^{2+} channel (N-type)	Nervous, neuroendocrine tissue
	$_{z}\alpha$	-	Adenylyl cyclase (-)	
	$_{t}1\alpha$	-	cGMP PDE	Rod
	$_{t}2\alpha$	-	cGMP PDE	Cone
	Gustducin	-	-	Taste buds
G_q	$_{q}\alpha$	42kD	PLCβ	Ubiquitous
	$_{11}\alpha$	43kD	PLCβ	Ubiquitous
	$_{14}\alpha$	-	-	Spleen, lung, kidney, testis, haematopoietic tissue
	$_{15}\alpha$	-	-	Haematopoietic cell types
G_{12}	$_{12}\alpha$	-	PLA_2	Ubiquitous
	$_{13}\alpha$		Na^+/H^+ exchanger	Ubiquitous

Table 3.1. Classification of the G α-subunit proteins. MW: molecular weight; PLC = phospholipase C; PLA_2 = phospholipase A_2; PDE = phosphodiesterase (modified from Schwindinger and Levine, 1997).

Different types of receptors are linked to different G proteins, for example β_1- and β_2- adrenergic receptors are coupled to different G_s proteins that activate adenylyl cyclase. In contrast, α_1 and α_2-adrenergic receptors are coupled to G_q and G_i, respectively. G_i inhibits adenylyl cyclase and G_q stimulates phospholipase C to generate second messengers (Schwindinger and Levine, 1997, Lodish et al., 2000).

$G_z\alpha$ has been purified from bovine brain and shows a slow rate of guanine nucleotide exchange when compared to $G_s\alpha$ and other $G_i\alpha$ proteins (Simon et al., 1991). The $G_o\alpha$ protein is the most abundant G protein in neurons, but is otherwise expressed only in endocrine cells and heart. The effects that the disruption of $G_o\alpha$ occasioned in transgenic mice demonstrated that this protein plays a major role in motor control, motor behaviour and in pain perception (Jiang M et al., 1998).

The genes encoding $G_{i1}\alpha$, $G_{i2}\alpha$, $G_{i3}\alpha$ and $G_o\alpha$ and $G\alpha$-subunits of the transducin rod and cone receptors all have similar structures with conserved positions of splice junctions. Several $G\alpha$ proteins are expressed ubiquitously, but this is not the case, for example, for the G_q group members: $G_{14}\alpha$ is expressed only in bone marrow cells, some early myeloid cells and in progenitor B cells; $G_{15}\alpha$ is also primarily expressed in hematopoietic cell types (Wilkie et al., 1991).

The $G_{12}\alpha$ subfamily has only two members, $G_{12}\alpha$ and $G_{13}\alpha$ that participate in cell transformation and embryonic development. $G_{13}\alpha$ participates in the regulation of cell movement in response to specific ligands, as well as in developmental angiogenesis. An activated form of $G_{13}\alpha$ induces cytoskeletal changes by way of Rho (Offermanns et al., 1997). The activities of members of the Rho family are regulated by guanine nucleotide exchange factors (GEFs). p115, a GEF specific for Rho, has an NH_2-terminal region with similarity to the conserved domain of the family of regulators of G protein signalling (RGS). Members of the RGS family can stimulate the GTPase activity of the α subunit of some G proteins. It has been demonstrated that p115 has specific activity as a GTPase activating protein toward the α subunits of $G_{12}\alpha$ and $G_{13}\alpha$, but not toward other G proteins. This GEF may act

as an intermediary in the regulation of Rho proteins by $G_{12}\alpha$ and $G_{13}\alpha$. It may define a subset of RGS proteins that have GAP activities but also couple RhoGEF activity to G proteins α subunits (Kozasa et al., 1998).

Gα-subunits have a role also in the regulation of the structure of the Golgi apparatus. The Golgi apparatus is involved in intracellular trafficking, and its structure can be affected by transport inhibitors such as nordihydroguaiaretic acid (NDGA) that cause Golgi disassembly. By over-expressing different Gα-subunits, Yamaguchi et al. (2000) demonstrated that at least in the case of the constitutively active form of $G_z\alpha$ and $G_{i2}\alpha$, such overexpression suppresses NDGA-mediated Golgi disassembly, indicating that these G proteins have a function in protecting the Golgi structure.

3.1.2.2.1.3 The Gβγ-subunit of heterotrimeric G proteins

Gβ and Gγ-subunits form a structural dimer that acts as a single entity. Gβ and Gγ chains are polypeptides of 35 000 and 8 000 D respectively (OMIM 2000). So far 6 genes that code for Gβ-subunits and 12 genes that code for Gγ-subunits have been identified (Miric et al., 1993, Schwindinger and Levine, 1997, Farfel et al., 1999, Lodish et al., 2000).

The Gβγ-subunit can facilitate Gα-subunit-receptor interaction, modulate effector activity or mediate receptor desensitivity (Lodish et al., 2000). The Gβγ complex can also interact directly or indirectly with different effectors (Kleuss et al., 1992). Despite their close relatedness, the different types of Gβ-subunit proteins cannot form dimers with all kinds of different Gγ-subunit. For example, $G\beta_1$ can associate with $G\gamma_1$ or $G\gamma_2$, whereas $G\beta_2$ can associate with $G\gamma_2$ but not with $G\gamma_1$, and $G\beta_3$ is unable to associate with either $G\gamma_1$ or $G\gamma_2$ (Clapham and Neer, 1993).

The five known mammalian β subunits are between 53% and 90% identical to each other. In contrast, the six γ subunits are much more different from each other than are the β subunits or the α subunits (Neer, 1995). The analysis carried out by

Rahmatullah and Robishaw (1994) and Rahmatullah et al. (1995) showed that selective interaction of particular Gα and Gγ-subunits provides the mechanistic basis for the assembly of specific Gα, Gβ and Gγ-subunits combinations. This association is dependent on prenylation of the Gγ-subunit.

3.1.2.2.2 Gα-subunit crystallographic studies

The conformation of the Gα-subunit of the activated form of $G_s\alpha$ was determined at 2.5 A resolution (Sunahara et al., 1997). $G_s\alpha$ (as is apparently also the case with other Gα-subunits) consists of two domains: a Ras-like domain common to all members of the GTPase superfamily (encoding the structural elements for guanine nucleotide binding), joined through two linker polypeptides to an α-helical domain unique to the family of heterotrimeric G proteins.

At the core of every G protein is the guanine nucleotide-binding domain, consisting of 200 residues with a central six-stranded β-sheet and five α-helices. Five of the six β-sheets (β_1-β_6) are parallel and one (β_2) is antiparallel to the others. The five α-helices (α_1-α_5) surround the β-sheets.

The helical domain has an entirely α-helical secondary structure with one long central helix (α_A) surrounded by five short helices (α_B-α_F). It is linked to the GTPase domain by two extended strands, linker 1 and linker 2. Between these two domains lies a deep cleft within which the nucleotide is tightly bounded.

GTP is the organizing centre for three switch elements (switch I, II and III) that undergo substantial conformational rearrangements on GTP hydrolysis. Apparently the structure and orientation of the switch elements are essentially identical in $G_s\alpha$, $G_i\alpha$ and $G_t\alpha$ (Sunahara et al., 1997). As described for $G_t\alpha$, switch I (Ser^{173}-Thr^{183}) encompasses linker 2 and extends partially into the β strand. Switch II (Phe^{195}-Thr^{215}) begins near the C-terminus of β_3, extends through α_2 and includes the α_2 and β_4 loop. Switch III (Asp^{227}-Arg^{238}) spans the β_4-α_3 loop and occurs within a sequence

segment that is the second of 4 sequence inserts (I1-I4) that are not present in the Ras sequence used for comparison (Lambright et al., 1994, Sprang, 1997). Switch I and II (at least in $G_t\alpha$) are analogous to switch regions in Ras but switch III is unique to the heterotrimeric family of G proteins. The movement of switches I and II upon GTP binding is in direct response to the presence of the γ-phosphate group. Switch III, however, has no direct contact with the bound guanine nucleotide. Upon activation, switches II and III move towards each other and form multiple interactions.

The analysis of mutations found in some PHP-Ia patients involving Arg^{258} and Gly^{259} residues (in $G_s\alpha$), suggests that switch III is involved in both the activating mechanism and in maintaining the basal state of this protein. Arg^{258} mutations have a direct result of decreased GDP binding due to loss of contacts between Arg^{258} side chain and residues within the helical domain. Gly^{259} interacts with conserved domains in switch II, a region that is important in maintaining the active state (Warner et al., 1998 and 1999).

3.1.2.3 The (effector) adenylyl cyclase enzyme

The third element of the signalling unit that interacts with $G_s\alpha$ is the effector adenylyl cyclase, whose activity is regulated by the activity of the Gα-subunit. The adenylyl cyclase enzyme catalyses the conversion of intracellular ATP to cAMP. At least 8 different forms of adenylyl cyclase (I-VIII) have been described. They have molecular weights of ~120 000 (range of 1080-1248 amino acid residues). The enzyme structure crosses the plasma membrane 12 times in 2 cassettes of 6 "transmembrane-spanning domains". The overall amino acid sequence similarity among the different isoforms of adenylyl cyclase is ~50% (Cooper et al., 1995, Taussig and Gilman, 1995). Most of the adenylyl cyclases are subject to multiple forms of regulation and besides the Gα-subunits they can also be stimulated or inhibited by, for example, protein kinase C, Ca^{2+} and Gβγ-subunits (Cooper et al., 1995).

3.1.3 MECHANISM OF ACTION OF THE G PROTEINS

As has been mentioned, the G proteins are turned "on" when bound to GTP and turned "off" when bound to GDP. In the basal state, in the absence of hormonal simulation, the protein is bound to GDP. Hormonal activation of the cell-surface receptor then activates the release of GDP; the subsequent binding to GTP over GDP is favoured by the higher concentrations of GTP in the cell. The intrinsic GTPase activity of these GTP-binding proteins hydrolyses the bound GTP to GDP and P_i, thus converting the active form back to the inactive form (Lodish et al., 2000).

Binding of many hormones and neurotransmitters to specific receptors stimulates activity of adenylyl cyclase and results in the production of the intracellular second messenger cAMP. The latter mediates the ultimate physiological effects of these agonists, through a phosphorylation cascade that is initiated by activation of protein kinase A. The mechanism of action of stimulatory G proteins can thus be summarized as (Figure 3.1):

1. Binding of hormone produces conformational change in the receptor.
2. The receptor binds to $G_s\alpha$.
3. Binding to receptor induces a conformational change in $G_s\alpha$. GDP bound to $G_s\alpha$ is replaced by GTP and the α subunit dissociates from $G\beta\gamma$-subunits
4. $G_s\alpha$ binds to adenylyl cyclase, activating synthesis of cAMP
5. Hydrolysis of GTP to GDP causes $G_s\alpha$ to dissociate from adenylyl cyclase and bind once again to $G\beta\gamma$ (Figure 3.1 modified from Lodish et al., 2000).

Figure 3.1. Mechanism of action of the G proteins. The steps are numbered 1 to 5 as explained in the text.

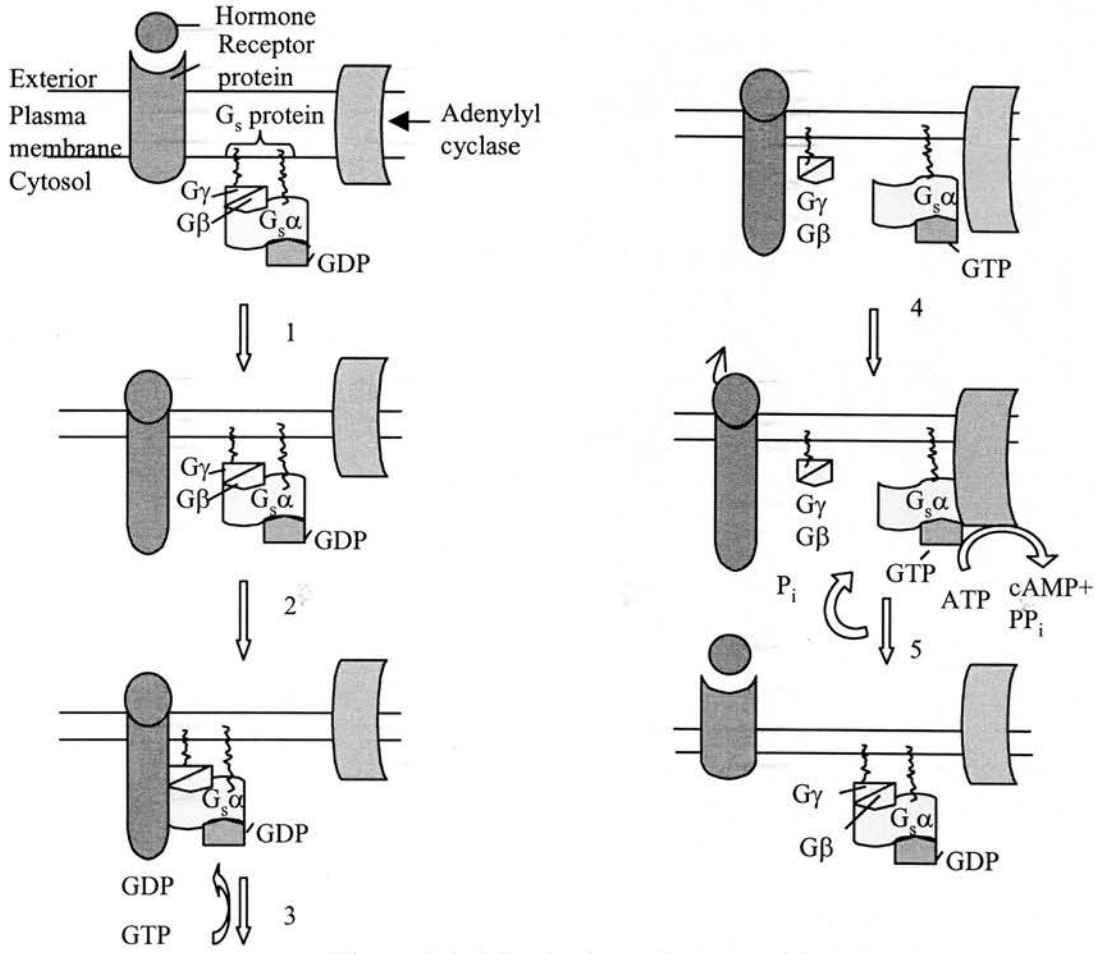


Figure 3.1. Mechanism of action of the G proteins

3.1.4 FACTORS MODIFYING G PROTEIN ACTIVITY

There are various factors that may modify the mechanism of action of G proteins. For example, it has been proposed that the Gβγ-subunit can play a role in communicating the activating signal from the receptor to the nucleotide-binding site, allowing the receptor to have effects on the α subunit that are not mediated by direct contact (Iiri et al., 1994 and 1998).

The rate of hydrolysis of GTP can also be modified, for example by substitution of the alanine residue at position 366 of the Gα-subunit. It has been observed that when this residue is mutated there is a spontaneous release of GDP at a rate 80 times faster than in the wild-type subunit (Iiri et al., 1994 and 1998).

Proteins that regulate the slow intrinsic rate of GTP hydrolysis by Gα-subunits are called GTPase-activating proteins or GAPs. One of the GAP members is the family known as regulators of G protein signalling or RGS proteins. Nineteen genes that code for these proteins have been described in mammals. RGS proteins are recognizable in sequence by the presence of a conserved 120 amino acid core region, which is flanked by variable arms, making up a protein of some 25kD overall (Berman and Gilman, 1998).

The RGS proteins are negative regulators of G protein signalling that act at the level of the dissociated Gα-subunit. Their activity appears to be influenced by palmitoylation of the Gα-subunit, as this modification decreases its affinity for certain RGS proteins (Berman and Gilman, 1998). Activation of G_s by the β-adrenergic receptor is accompanied by an increase in the turnover of palmitate on the $G_s\alpha$ subunit (Mumby, 1997). Although the RGS proteins had no effect on the time course of nucleotide binding, they did potently stimulate (by at least 40-fold) the rate of GTP hydrolysis by several G proteins (Koelle, 1997) (Figure 3.2), indicating that RGS proteins downregulate G protein action by virtue of their GAP activity.

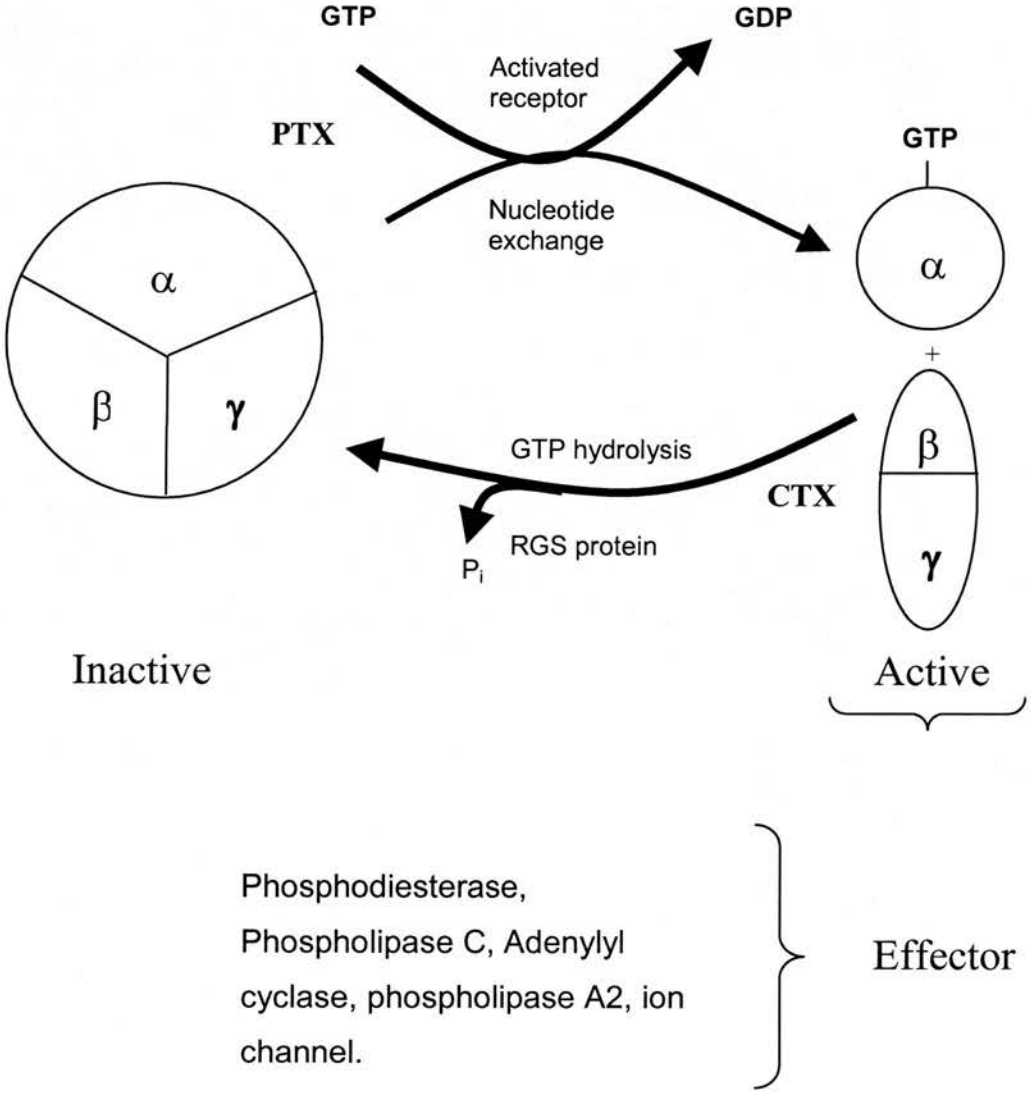


Figure 3.2. GTPase cycle for heterotrimeric G proteins. The position in the cycle of the RGS proteins is shown. The places in the cycle where Pertussis toxin (PTX) and Cholera toxin (CTX) can exert their action are indicated. P_i, phosphate (modified from Koelle, 1997 and Hadley, 2000).

Changes in the regulation of GTPases that have an important pathological repercussion are those caused by pertussis toxin (PTX) and cholera toxin (CTX). Pertussis toxin (secreted by *Bordetella pertussis*) ADP-ribosylates a cysteine residue located in the C-terminal tails of the Gα-subunit of the $G_i\alpha$ family. This has the effect of blocking the catalysis of GTP exchange by the activated receptor, leaving the Gα-subunit unable to interact with a receptor or to exchange GDP for GTP, and locking both Gβγ and Gα-subunits in their inactive states (Clapham and Neer, 1993, Hadley, 2000) (Figure 3.2).

The cholera toxin (produced by the bacterium *Vibrio cholerae*) is an hexameric protein containing 1 α subunit and 5 β subunits. The α subunit is an enzyme that penetrates the plasma membrane and enters the cytosol, where it catalyses the covalent addition of the ADP-ribose moiety from intracellular NAD+ to $G_s\alpha$ at the arginine reside at position 201. ADP-ribosylated $G_s\alpha$-GTP can activate adenylyl cyclase normally but cannot hydrolyse the bound GTP to GDP; thus $G_s\alpha$ remains in the active state, continuously activating adenylyl cyclase (Spiegel et al., 1985, Farfel et al., 1999, Lodish et al., 2000, Hadley, 2000) (Figure 3.2). These two agents have been useful to characterise the mechanism of action of specific heterotrimeric G proteins depending upon how they are affected by treatment with one or other toxin.

3.1.5 THE *GNAS1* GENE

Bray et al. (1986) described four species of $G_s\alpha$ cDNA in humans; they were named $G_s\alpha$-1 to 4. One of the clones described in their study ($G_s\alpha$-1) had an open reading frame of 1152 nt and coded for a protein of 394 amino acid residues. Its sequence was 95% homologous to the bovine $G_s\alpha$ sequence used for comparison. The four human $G_s\alpha$ sequences differed from each other in a region that corresponded to amino acid residues 71-87 of the bovine $G_s\alpha$ gene.

As mentioned, $G_s\alpha$-1 closely resembles bovine $G_s\alpha$; $G_s\alpha$-2 differs from $G_s\alpha$-1 in codon 86 (Gly) and the presence of three additional nucleotide residues (AGT) that

code for Ser^{87}. $G_s\alpha$-3 has the codon GAT for Asp^{71} instead of GAG for Glu^{71} and lacks codons for amino acid residues 72-86 of $G_s\alpha$-1. $G_s\alpha$-4 has a GAC rather than a GAT codon for Asp^{71}, lacks codons for amino acid residues 72-86 and contains an AGT codon for Ser^{72}.

An alternative splicing scheme was proposed to explain how these four classes of mRNA were derived from the same gene (later verified by Kozasa et al. (1988)). The relative abundance of the different types of $G_s\alpha$ mRNAs (as assessed in human basal ganglia) were 33%, 11%, 23% and 3% for $G_s\alpha$-1 to 4 respectively. The $G_s\alpha$-1,2 and $G_s\alpha$-3,4 isoforms correspond to the 52 kD and 45 kD forms of $G_s\alpha$ protein.

The genomic structure of the gene that codes for $G_s\alpha$, *GNAS1* was reported by Kozasa et al. (1988). The genomic structure described was composed of 13 exons and 12 introns, spanning a region of ~20 kb. All the splice junctions conformed to the GT-AG rule. The information obtained provided an explanation for the basis for the alternative splicing that generates four isoforms of $G_s\alpha$ mRNA.

Isoform $G_s\alpha$-3 lacks a stretch of 45 nt present in $G_s\alpha$-1, which coincides with exon 3. Therefore $G_s\alpha$-1 and $G_s\alpha$-3 are derived from the alternative splicing (by "exon skipping") of exon 3. $G_s\alpha$-2 has three nucleotides additional to $G_s\alpha$-1, inserted at the 3′ end of exon 3. $G_s\alpha$-4 also has these three nucleotides additional to $G_s\alpha$-3, in this case between exons 2 and 4. These four types of $G_s\alpha$ mRNA are thus generated from a single gene by alternative use of exon 3 and/or two alternative splice acceptor sites of intron 3. These are only three nucleotides apart; at the 3′ end of intron 3 is the unusual splice junction sequence CTGCAG, which can be spliced in either of the patterns CTG\CAG or CTGCAG\. In the former case, use of the non-consensus CTG instead of the consensus CAG results in the insertion of a single Ser codon in the $G_s\alpha$-2 and $G_s\alpha$-4 mRNAs (Kozasa et al., 1988, Schwindinger and Levine, 1997).

The GC content of the 5′ region (including and upstream of exon 1) is very high, 85%. An S1 nuclease mapping analysis showed multiple transcriptional initiation sites. Comparing this original description of the *GNAS1* genomic structure to the organisation of the human $G_i\alpha$ gene revealed common features. The two genes have

three identical exon junctions: the exon 1-2, 6-7 and 9-10 junctions of $G_s\alpha$ correspond precisely to the junctions of $G_i\alpha$ exons 1-2, 4-5 and 5-6 respectively. This suggests that Gα subunits of signal transducing proteins evolved from a common ancestral gene (Kozasa et al., 1988).

The mapping of *GNAS1* to human chromosome 20 was first performed by using a cDNA probe to screen a mouse/human somatic cell hybrid panel. The gene was later located to 20q13.2-q13.3 by *in situ* hybridization (Sparkes et al.,1987, Levine et al., 1991).

Subsequent reports demonstrated that the genomic structure originally reported for *GNAS1* was only partial, as additional exons were identified. A $G_s\alpha$ mRNA that uses a different promoter and leading exon that lies ~2.5 kb 5′ of the original exon 1 (exon A), does not contribute an in-frame ATG. Its mRNA is therefore either not translated, or less likely, encodes an N-terminally truncated form of $G_s\alpha$ by use of a non-consensus downstream start codon (Ishikawa et al., 1990, Swaroop et al., 1991).

Another cluster of novel *GNAS1* exons was defined much farther upstream, during the studies described in this chapter. This work was initiated by the results of a project undertaken by our group in collaboration with the laboratory of Y. Hayashizaki. A screen for novel imprinted loci was carried out by restriction landmark genomic scanning for methylation differences (RLGS-M) between normal and parthenogenetic human DNA. One differentially methylated spot, designated A20, was identified, subsequent analysis of which showed it to contain an exon that was part of the *GNAS1* gene. Further analysis of the upstream genomic region identified other exons besides A20 and showed that the *GNAS1* genomic structure is much more complex and spans a bigger chromosomal region that originally thought (see Discussion).

One transcript identified in these studies encodes a larger $G_s\alpha$ isoform named XLαs (extra-large $G_s\alpha$-related protein). In this mRNA, a single large novel upstream exon is spliced to exons 2-13 of $G_s\alpha$. This yields a long open reading frame continuous

with that of exons 2-13, encoding a protein with a large novel N-terminal domain in place of the exon 1-encoded N terminus of $G_s\alpha$. XLαs was originally identified in the rat while searching for proteins that undergo ADP-ribosylation by cholera toxin *in vivo*. The strategy used was to incubate PC12 cells in the absence or presence of cholera toxin and then subject a postnuclear supernatant prepared from these cells to "back-ribosylation" *in vitro* using cholera toxin and ^{32}P-labelled NAD. A protein that had become ADP-ribosylated by cholera toxin *in vivo* will thus show a reduced incorporation of ^{32}P ADP-ribose in the subsequent *in vitro* reaction. A protein with apparent M_r of 94 kD, referred to as XLαs (for extra large αs) was identified.

The original description of XLαs (Kehlenbach et al., 1994) was of a cDNA (2956 nt) containing an open reading frame of 2538 nt, beginning with an ATG at nt 46-48 and encoding an 846 amino acid residue 92 kD protein. However, later analyses indicated that the translation of XLαs probably actually starts at nt 439-441, resulting in a 715 amino acid residue 78 kD protein (this protein migrates anomalously on SDS-PAGE, with an apparent mobility of 94kD, which was the reason for the original erroneous assignment of the upstream ATG as the start codon). The XLαs protein consists of two portions: an N-terminal portion (referred as the "XL portion" of 367 amino acids) and a C terminal portion (the "αs portion") that is identical to amino acid residues 47-394 of $G_s\alpha$. The XL portion replaces the portion of $G_s\alpha$ encoded by exon 1. The XL portion of rat XLαs appears to contain several structurally distinct regions; an acidic polar N-terminal domain followed by a very alanine rich domain containing a tetrapeptide repeat based on EPAA repeats, a proline-rich hinge, a cysteine-rich domain and finally a domain with homology to exon 1 as in $G_s\alpha$ (Kehlenbach et al., 1994 and 1995).

The predicted human protein has a tripeptide repeat motif, which is again rich in proline, alanine, and an acidic amino-acid (aspartate rather than glutamate) (Hayward et al., 1998a). XLαs is localized in the plasma membrane, and like $G_s\alpha$, can bind G protein $\beta\gamma$ subunits and activate adenylate cyclase (Klemke et al., 2000, Pasolli, et al., 2000). However, a transmembrane receptor able to activate XLαs has not been identified.

3.1.5.1 Gnas gene in mouse

From the studies carried out by Cattanach and Kirk (1985), using chromosomal translocations in mice to generate offspring with segmental uniparental disomy (UPD), it was demonstrated that for many chromosomal regions, proper development in mammals requires both maternal and paternal contributions indicating the presence of imprinted genes. In these studies, the distal part of mouse chromosome 2 (a region of conserved synteny corresponding to human 20q) was recognised as being imprinted. Because two different phenotypes resulted from maternal and paternal UPD, the presence of both maternally and paternally imprinted genes in this region was initially suspected. Later analyses have shown that within this region of mouse chromosome 2 two different regions subject to genomic imprinting can be identified. Molecular analyses have confirmed the conclusions of these classical mouse genetics studies, since imprinted genes have been identified in both regions by applying cDNA subtractive hybridisation and differential screening methods (Kagitani et al., 1997, Kikyo et al., 1997, Williamson et al., 1998a).

Ashley et al. (1987) mapped the *Gnas* gene in the mouse to chromosome 2, which, by the argument of homology of synteny, suggested an assignment of human *GNAS1* to human chromosome 20 (confirmed by Sparkes et al., 1987). The mouse gene was later mapped to a 20 centimorgan region between breakpoints T2Wa and T28H on distal chromosome 2, a region of conserved synteny to human 20q13-14 (Williamson et al., 1996).

The finding that *Gnas* was located in an imprinted chromosomal region prompted studies to determine if this gene was indeed under this form of regulation. The knowledge that the $G_s\alpha$ protein (encoded by the human homologue *GNAS1*) was deficient in patients suffering from PHP-Ia, whose full clinical presentation seemed to depend on maternal inheritance of the mutation, seemed to indicate that the human gene, too, might be imprinted.

A study of *Gnas* expression by *in situ* hybridization to tissues of mice carrying maternal duplication/paternal deficiency for distal Chr2 (MatDp2) and its reciprocal (PatDp2) was carried out by Williamson et al. (1996). Unexpectedly, however, they described imprinting in mice in a manner opposite to that predicted by observations in AHO. The RNA *in situ* hybridisation revealed high levels of *Gnas* mRNA in glomeruli of PatDp2 embryos at late gestation and much lower levels in glomeruli of MatDp2 embryos. This implied that the maternal, rather than the paternal, *Gnas* allele is silenced in renal glomeruli.

Different conclusions were reached by Yu et al. (1998), who generated mice with a null *Gnas* allele resulting from disruption of exon 2. This demonstrated, firstly, that homozygous Gs deficiency is embryonically lethal. Maternal ($Gnas^{m-/p+}$) and paternal ($Gnas^{m+/p-}$) heterozygotes for the null allele were shown to have distinct phenotypes, suggesting both maternal and paternal imprinting. Resistance to parathyroid hormone was present in *Gnas* m-/p+ but not in *Gnas* m+/p- mice, suggesting monoallelic $G_s\alpha$ expression in the PTH target tissues (proximal renal cortex). In confirmation of this, in renal cortex (and in adipose tissue) $G_s\alpha$ protein and mRNA levels were much lower in m-/p+ than in m+/p- mice. These findings support the conclusion that *Gnas* is imprinted in mice and suggest that similar imprinting in human *GNAS1* may explain the variable phenotypes observed in PHP-Ia.

The uniparental disomy phenotypes described by Cattanach and Kirk (1985) were studied *in utero* by Williamson et al. (1998b). They generated mice and embryos with maternal duplication/paternal deficiency (MatDp.dist2) and the reciprocal (PatDp.dist2) of chromosome 2. They examined the uterine contents at several stages and carried out weight comparisons between Pat and MatDp.dist2 mice and their normal sibs. The weights were measured directly (wet weight) and also after 72 hours in a freeze-dryer (dry weight).

The PatDp.dist2 mice had an increase in wet weight compared to normal mice, but this had diminished by birth and the placental weight was slightly diminished in the second half of gestation. It was also noted that at 16.5 dpc, PatDp.dist2 embryos

were severely oedematous and that the newborns had an increased length of the long bones. On the other hand, the wet weight ratios of MatDp.dist2 embryos decreased during the second half of the gestation, but placental weight was normal. The dry weight analysis indicated that (at 16.5 dpc) there was no difference in either PatDp.dist2 or MatDp.dist2 compared with normal, so the previous differences observed were due to fluid retention. The histological analysis of several internal organs revealed no overt abnormalities in both types of mice.

3.1.6 MUTATIONS OF G PROTEINS

When the ability of a G protein to switch to the "on" state is altered, this change may result in an increase or a decrease in the downstream signal. If the signal is increased it results in the defective protein releasing GDP and binding GTP more rapidly than normal; conversely a decreased signal results when the protein releases GDP more slowly or binds GTP less avidly (Farfel et al., 1999).

A defect of G protein-coupled signal transduction could result from quantitative and/or qualitative changes in GPCRs and G proteins. Such changes can result from mutations in the genes that code for both kinds of structure; the phenotypic result will depend upon the biological activities and range of expression of the abnormal gene. Factors relating to the mutation itself include whether it occurred in the germ line or was somatic and also the precise nature of the mutation, resulting in a gain or loss-of function of the protein involved (Spiegel, 1996).

3.1.6.1 Gain-of-function mutations in G proteins

Gain-of-function mutations lead to inappropriate activation of a G protein. Constitutive activation will result in effector signalling despite the lack of normal upstream signals from a hormone-activated GPCR. Such mutations act in one of two ways: they can either accelerate the release of GDP and cause receptor-independent G protein activation, or block the GTPase reaction that terminates G protein activation. The effects of the mutation are likely to be dominant, causing disease in

heterozygous individuals, and may mimic the effects of hypersecretion of the hormone normally activating the involved GPCR or G protein. However, the circulating concentration of the hormone will actually be suppressed, reflecting autonomous hyperfunction of the target gland (Spiegel, 1996, Cohen and Howell, 1999) and physiological feedback inhibition of hormone secretion.

At the receptor level, gain-of-function mutations have been identified, for example in the receptors for LH and TSH. In the former case they have been found in the germline in cases of familial male precocious puberty, and in the case of the TSH receptor, somatically acquired gain-of-function mutations have been described in hyperfunctioning thyroid adenomas (Spiegel, 1996).

3.1.6.2 Loss-of-function mutations in G proteins

Loss-of-function mutations can operate at various levels, ranging from blocking normal mRNA and/or protein synthesis, to preventing the synthesized protein from reaching its normal subcellular location, to impairing biochemical function despite synthesis and normal targeting of the protein. Loss of function mutations in G proteins may block receptor coupling or activation by GTP (Spiegel, 1996, Schwindinger and Levine, 1997). Some loss-of-function mutations behave dominantly and cause clinical disease in heterozygotes. Loss-of-function of a given GPCR or of the G protein to which it is coupled will cause hormone resistance, with a clinical phenotype that will resemble that resulting from deficiency of the hormone itself. However, it is likely to be associated (conversely to the gain-of-function mutation effect) with a high circulating concentration of the corresponding hormone agonist (Spiegel, 1996).

3.1.6.3 Clinical conditions caused by mutations in the Gα-subunit

Several conditions caused by mutations in different members of the four groups of G protein α-subunits have been described, and are listed in Table 3.2.

CAUSE AND DISEASE	PROTEIN	MOLECULAR MECHANISM	DISTRIBUTION
Caused by defective signal termination - signal excessive.			
Cholera	$G_s\alpha$	ADP-ribosylation of ARG^{201} inhibits GTP hydrolysis	Intestinal epithelium
Adenomas of the pituitary and thyroid	$G_s\alpha$	Point mutation Arg^{201} or Gln^{227} inhibits GTP hydrolysis.	Sporadic (somatic mutation)
Adenomas of the adrenal and ovary	$G_{i2}\alpha$	Point mutation Arg^{179} inhibits GTP hydrolysis.	Sporadic (somatic mutation)
McCune-Albright syndrome	$G_s\alpha$	Point mutation Arg^{201} inhibits GTP hydrolysis	Mosaic (mutation in early embryo)
Caused by absent or inactive $G\alpha$ - signal deficient.			
Pseudohypo-parathyroidism type Ia	$G_s\alpha$	One null $G_s\alpha$ allele decreases response to hormones parathyroid hormone, thyrotropin, gonadotropins	Germ-line mutation
Type Ib	$G_s\alpha$	The sole defect is decreased response to parathyroid hormone	Germ-line mutation
Night blindness	$G_t\alpha$	Point mutation Gly38Asp, mechanism is not known.	Germ-line mutation
Caused by abnormal signal initiation - signal inadequate or excessive.			
Pertussis	$G_i\alpha$	ADP-ribosylation of a cysteine blocks activation by receptor and decreases signal	Bronchial epithelium
Pseudohypo-parathyroidism type Ia	$G_s\alpha$	Point mutation Arg385His blocks activation by receptor and decreases signal. Point mutation Arg231His impairs GTP binding and decreases signal	Germ-line mutation
Testotoxicosis with pseudohypo-parathyroidism type Ia	$G_s\alpha$	Point mutation Ala366Ser accelerates GDP release, enhances signal at 34°C (testis) but destabilizes and inactivates $G_s\alpha$ at 37° (pseudohypoparathyroidism type Ia)	Germ-line mutation.

Table 3.2. Clinical conditions associated with Gα-subunit mutations (modified from Farfel et al., 1999).

As can be seen in Table 3.2, the clinical and physiological consequences of the different mutations in α-subunits are wide and several of them have been characterised in detail. Among them is one form of dominantly inherited congenital night blindness ("Nougaret" type) that is caused by a missense mutation in the gene encoding the Gα-subunit of rod transducin, the G protein that couples rhodopsin to cGMP-phosphodiesterase in the phototransduction cascade. The mutation, identified in exon 2, corresponded to the change Gly38Asp. Gly^{38} is a highly conserved residue among heterotrimeric G proteins ($G_z\alpha$ is the only known G protein without a glycine at this position). The missense mutation altering the homologous residue Gly12Val was the first oncogenic mutation discovered in $p21^{ras}$, and it may interfere with the hydrogen bonds to the phosphate groups of GTP/GDP (Dryja et al., 1996).

Before discussing $G_s\alpha$ mutations that have been implicated in single gene disorders it is also worth noting that an association study of *GNAS1* has also been carried out to assess its possible role in essential hypertension (Jia et al., 1999). By using the *Fok*I polymorphism in exon 5, the authors demonstrated a significant difference in frequency of the *Fok*I alleles between 268 white hypertensive individuals and a matched group of 231 control subjects. These results suggested that the *GNAS1* locus might carry a functional variant that influences blood pressure variation and response to β-blockade in essential hypertension, perhaps causing increased cAMP-mediated cardiac contractility (Jia et al., 1999).

$G_s\alpha$ is not the only G protein subunit that has been associated with essential hypertension. A polymorphism (C825T) in exon 10 of the gene encoding the $G\beta_3$-subunit of heterotrimeric G proteins was identified in some patients with essential hypertension (Siffert et al., 1998). The polymorphism does not affect the amino-acid sequence of Gβ3, but in individuals carrying the T allele it was associated with the occurrence of a splice variant (*GNB3*-s) in which nucleotides 498-620 of exon 9 are deleted, causing the in-frame loss of 41 amino acids and one of the so-called WD repeat domains of the Gβ subunit. This alternative splicing results from the activation of a cryptic splice acceptor site, and the resulting Gβ3-s protein appears to be a biologically active variant.

Genotype analysis of 427 normotensive and 426 hypertensive subjects suggested a significant association of the mutated allele with essential hypertension. Although the presence of Gβ3 in cultured cells correlates well with increased G_i-dependent hormone responses, the underlying molecular mechanism is not known yet (Iiri and Bourne, 1998, Siffert et al., 1998, Farfel, 1999).

3.1.7 MUTATIONS IDENTIFIED IN *GNAS1*

Both gain and loss-of-function mutations have been described in the *GNAS1* gene. As has been explained before, the mutations can result in two different effects. Mutations in *GNAS1* that impair the ability of $G_s\alpha$ to hydrolyse GTP lead to constitutive activation of adenylyl cyclase and result in hyperfunction and autonomous growth of endocrine tissues. By contrast, mutations in *GNAS1* that lead to decreased expression or function of $G_s\alpha$ are associated with generalised resistance to hormones that act by stimulating adenylyl cyclase (Schwindinger and Levine, 1997, Cohen and Howell, 1999).

3.1.7.1 Gain-of-function mutations of GNAS1

Gain-of-function mutations of *GNAS1* have been identified in McCune-Albright syndrome (MAS) (OMIM#174800), polyostotic and monostotic fibrous dysplasia and in certain endocrine tumours, notably pituitary somatotroph adenomas.

3.1.7.1.1 McCune-Albright (MAS) syndrome

McCune-Albright syndrome (OMIM #174800) is a sporadic condition, characterised by the clinical triad of polyostotic fibrous dysplasia, café-au-lait pigmented skin lesions and endocrine dysfunction. The latter can be manifested as precocious puberty, autonomous thyroid nodules, growth hormone excess and hyperprolactinemia. The patients develop single or multiple skeletal lesions that can

cause fracture, deformity and nerve entrapment (Albright et al., 1937, Wilson and Foster, 1992).

From the observation that MAS occurs sporadically, and from the variable, patchy distribution of the lesions, the suggestion was made early on that the disorder was caused by somatic mosaicism, for a mutation that would be lethal in non-mosaic state. Somatic mosaicism takes place when a mutation occurs postzygotically in a somatic rather than in a germ cell. One would postulate that all cells that descend from the mutated cell will manifest MAS related characteristics, while cells descended from non-mutated cells will develop into normal tissues, thus explaining the variability in tissue distribution. However, although the hypothesis that MAS results from somatic mosaicism for a mutation has been verified, recent molecular and cellular analyses suggest that the bone lesions of MAS are not in fact clonal, but contain a mixture of mutated and normal cells (Bianco et al., 1998).

Activating mutations of Arg^{201} have been found in a mosaic distribution in tissues from subjects with MAS. The mutation at codon 201, within exon 8 of *GNAS1*, is usually a C to T change resulting in Arg (CGT) to Cys (TGT). However, other mutations involving the same residue, Arg (CGT) to His (CAT) and Arg (CGT) to Gly (GGT) have also been described (Weinstein et al., 1991, Schwindinger and Levine, 1997). Arg^{201} mutations have also been described in patients with polyostotic fibrous dysplasia (a phenotypic variant of MAS) (Spiegel, 1996, Riminucci et al., 1999).

3.1.7.1.2 Oncogenic associations of GNAS1 (the gsp oncogene)

Vallar et al. (1987) described an altered $G_s\alpha$ protein in human growth hormone-secreting pituitary adenomas. These tumours were characterised by high secretory activity and intracellular cAMP levels. When cell membranes from these tumours were stimulated by growth hormone-releasing hormone, GTP or by fluoride, no effect upon adenylyl cyclase was observed. This finding indicated that the alteration of $G_s\alpha$ could be the direct cause of the high secretory activity of the tumours in

which it occurs. It represented the first description of an altered G protein in human disease.

The first oncogenic Gα-subunit mutants were found in the $G_s\alpha$ gene of pituitary tumours from patients with acromegaly. A role for $G_s\alpha$ in tumorigenesis was suggested by Landis et al. (1989), who established that the molecular basis for the previously demonstrated constitutive activation of adenylyl cyclase was a mutation in *GNAS1* that resulted in replacement of either Arg^{201} or Gln^{227}. Both of these residues are located in the GTP-binding site and are important for GTPase activity. Mutations of $G_s\alpha$ at Arg^{201} inhibit the GTPase activity and result in a constitutively active state, with ligand-independent stimulation of adenylyl cyclase (Spada et al., 1990). Gln^{227} in $G_s\alpha$ is analogous to the cognate amino acid Gln^{61} that is frequently replaced in the activated oncogene $p21^{ras}$; point mutations in the latter gene that inhibit GTPase activity are present in a variety of human tumours (Landis et al., 1989).

As the Arg^{201} and Gln^{227} mutations share some transforming characteristics with $p21^{ras}$, activated *GNAS1* has been referred to as the *gsp* oncogene (for G stimulatory protein). 40% of human pituitary somatotroph tumours have *gsp* mutations. The proteins encoded by *gsp* have the same defects in signal termination that occur in cholera: the inability to hydrolyse GTP and the persistent stimulation of adenylyl cyclase (Schwindinger and Levine, 1997, Farfel et al., 1999). The resulting marked elevation of intracellular cAMP results in the dominant character of the mutation (Clementi et al., 1990). Somatotroph tumours resulting from *gsp* mutations have different biological characteristics from those which are *gsp*-negative (Landis et al., 1989). They show greater sensitivity to inhibitory agents (somatostatin and its therapeutic analogue octreotide, and also dopamine). This sensitivity to inhibition may partly explain the low rate of tumour growth (Clementi et al., 1990, Spada et al., 1990).

Similar findings of mutations involving Arg^{201} and Gln^{227} have been reported in toxic thyroid adenomas, adrenocortical adenomas and parathyroid tumours (Yoshimoto et

al., 1993, Boothroyd et al., 1995). Thyroid-specific expression in transgenic mice of $G_s\alpha$ with the Arg^{201} to His mutation provided direct confirmation that the mutant form of $G_s\alpha$ has an oncogenic activity, supporting the model that deregulation of cAMP level alters growth control (at least in thyroid epithelium) (Michiels et al., 1994).

3.1.7.2 LOSS-OF-FUNCTION MUTATIONS OF *GNAS1*: PSEUDOHYPOPARATHYROIDISM

The first clinical and biochemical description of pseudohypoparathyroidism (PHP) was made in 1942 by Albright et al. They described three patients with apparent hypoparathyroidism (HP) but in whom after treatment with parathyroid extract no positive response was achieved, showing no phosphate diuresis or increase in phosphate clearance. Albright et al. (1942) initially used the term "Seabright-Bantam syndrome" for this entity (in reference to the abnormal feathering pattern of the male Seabright Bantam, that they believed resulted from failure to respond normally to testicular hormones). In this they recognized for the first time the existence of conditions in which they thought the primary disturbance was a failure of response to hormone(s) by the end-organ, rather than a failure in the nature of the hormone in itself (Albright et al., 1942, Wilson and Foster, 1992).

3.1.7.2.1 Clinical characteristics

Albright's clinical account of the phenotypic characteristics of these patients constitutes what is now known as Albright hereditary osteodystrophy (AHO) (OMIM 300800) and includes short stature, round face, cataract, nystagmus, short thick neck, obesity, reduced intelligence, seizures, hypocalcemic tetany, subcutaneous calcification or ossification and bone abnormalities such as osteoporosis, shortening of metacarpals and metatarsals. The Chvostek and Trousseau signs are also positive in these patients, reflecting hypocalcaemia (Albright et al., 1942, OMIM, 2000).

The clinical laboratory abnormalities seen are a high basal thyroid-stimulating hormone level and a hyper-response to thyrotropin-releasing hormone, low or absent antithyroid antibody titres, low serum calcium, high serum phosphate, and high levels of serum parathyroid hormone (PTH) (Laycok and Wise, 1996).

3.1.7.2.2 Molecular basis of PHP

Two related observations helped to elucidate the molecular basis for PHP. The first of them was the clinical observation that the administration of PTH to these patients failed to provoke a phosphate diuresis or an increase in serum calcium. The second observation was made by Chase et al. (1969) who demonstrated that the actions of PTH on its target organs (bone and kidney) were mediated by cAMP. The observation that PTH infusion leads to increased urinary excretion of nephrogenous cAMP proved to be a reliable diagnostic test for PTH resistance (Wilson and Foster, 1998).

In contrast to patients with idiopathic or postsurgical HP, patients with PHP show a markedly blunted urinary cAMP response to exogenous PTH, whereas patients with PPHP showed a normal response. Taken together, these results indicated that the biochemical defect in PHP was not in PTH itself but rather at or downstream of the receptor site in the plasma membrane (Wilson and Foster, 1992).

The biochemical defect in PHP patients was eventually shown to be a partially deficient activity of $G_s\alpha$. The activity of $G_s\alpha$ in two assay systems was significant lower (almost 50%) in erythrocyte membranes from PHP patients compared to normal subjects. The results supported the hypothesis that deficient activity of the guanine nucleotide regulatory protein was the molecular basis for hormone resistance in this inherited disorder (Farfel, et al., 1980 and 1981, Levine et al., 1980).

Levine et al. (1986) measured the $G_s\alpha$ activity of erythrocyte membranes both from AHO patients with clinical hormone resistance (PHP-Ia) and from family members with AHO alone (PPHP). Patients with PPHP had reductions in $G_s\alpha$ activity

comparable to those in patients with PHP-Ia, despite the fact that unlike patients with PHP-Ia, individuals with PPHP did not have obvious endocrine dysfunction. These findings generated the suggestion that more than one mechanism may be involved in the presentation of the disease (Levine et al., 1988).

Subjects with PHP and deficient $G_s\alpha$ activity show a high incidence of resistance to additional hormones (Wilson and Foster, 1992 and 1998). Normal cAMP response to PTH has been documented in osteoblasts isolated from patients with PHP-Ia and PHP-Ib, suggesting that the hypocalcaemia in PHP is not due to skeletal resistance, but is a consequence of the renal resistance to PTH, resulting in both increased urinary calcium losses and impaired 25 (OH) D_3 1α-hydroxylation (Wilson and Foster, 1998).

3.1.7.2.3 Classification of PHP

The classification of PHP is based on the clinical and biochemical characteristics of the patients. In PHP type I both the nephrogenous cAMP and phosphaturic response to PTH are blunted, whereas in PHP type II, there is a normal nephrogenous cAMP response to PTH infusion but a diminished phosphaturic response. These findings suggested that the defect in PHP type I was likely to be in the PTH receptor-adenylyl cyclase system and the defect in PHP type II was an inability to respond to cAMP.

PHP type I can be subclassified as follows:

1. Patients with PHP type Ia (PHP-Ia) have AHO and resistance to multiple hormones due to a deficiency of $G_s\alpha$ activity.

2. Patients with PHP type Ib lack AHO, have normal $G_s\alpha$ activity in erythrocytes and have hormone resistance that is limited to PTH. This has been presumed due to a defect in the PTH receptor, but recent studies in some families at least provide evidence for a tissue-restricted defect of $G_s\alpha$.

3. Patients with PHP type Ic have AHO and resistance to multiple hormones but $G_s\alpha$ activity is normal.

A diagnostic problem is found in patients presenting with overt hypocalcaemia and hyperphosphatemia, to discriminate between HP and PHP-I and II. For this purpose, responses to exogenous PTH-(1-34) (synthetic PTH) are a useful test. Data from HP patients provide the proper reference group, as the difference between HP and PHP is enhanced by the fact that the phosphaturic and cAMP excretory responses to PTH are both greater in HP patients than in normal subjects. PTH (1-34) is a safe and effective diagnostic agent for the investigation of renal responses to PTH. Patients with PHP have blunted cAMP and phosphaturic responses to PTH-(1-34) compared to those of either normal or HP subjects. Induced hypercalcaemia failed to restore a normal cAMP response to PTH-(1-34) infusion in 2 patients with PHP (Mallette et al., 1988).

Their different clinical presentation and endocrine abnormalities suggested that (at least) PHP-Ia and PHP-Ib were unrelated disorders and therefore thought to be caused by distinct molecular defects. The PTH-restricted nature of hormone resistance in PHP-Ib suggested that the defect might lie at the level of the specific receptor for PTH/PTH-related protein (PTHrP). However, in two studies of the receptor gene in patients with PHP-Ib, no mutations were found by direct sequencing of PCR products, suggesting that this was not the mechanism underlying PHP-Ib, at least in the patients analysed (Schipani et al., 1995, Fukumoto et al., 1996).

A genome-wide linkage scan using four PHP-Ib families with ~350 microsatellite markers mapped the PHP-Ib gene to the telomeric region of chromosome 20q13.3. Interestingly, in each kindred one or several individuals carried the allele associated with PHP-Ib but showed no biochemical evidence for the disorder; these unaffected gene carriers were always offspring of obligate male gene carriers. This suggested that PHP-Ib, like PHP-Ia, was paternally imprinted and maps to the chromosomal region containing *GNAS1* (Juppner et al., 1998). This unexpected finding further suggests that PHP-Ia and PHP-Ib (at least some cases) are after all allelic disorders.

It has recently been discovered that PHP-Ib patients often have an abnormality of *GNAS1* imprinting (see Discussion section 3.4.5). In one other recent report (Wu et al., 2001) the *GNAS1* mutation ΔIle^{382}, removing one residue from the C-terminus of $G_s\alpha$, was found in a PHP-Ib family. Co-transfection experiments showed that this mutation selectively impairs $G_s\alpha$ coupling to the PTH receptor, but not to the other $G_s\alpha$ -coupled receptors for TSH and LH, providing an explanation at the level of receptor-G protein interaction for the selective hormone resistance in PHP-Ib.

3.1.7.2.4 Molecular basis for hormone resistance in PHP

The observation that patients with AHO may have either reduced or normal levels of $G_s\alpha$ mRNA first suggested that $G_s\alpha$ deficiency might arise from a variety of genetic mutations. Several heterozygous *GNAS1* gene mutations that accounts for $G_s\alpha$ deficiency have been identified in the different families studied. These include an initiator codon mutation, missense mutations, mutations in sequences necessary for correct splicing and small deletions (Patten et al., 1990, Weinstein et al., 1990, Miric et al., 1993, Schwindinger et al., 1994, Farfel et al., 1996 and 1999, Schwindinger and Levine, 1997, Aldred and Trembath, 2000).

One of the most common mutations found in *GNAS1* in PHP-Ia is a 4 bp deletion in exon 7 (Yu et al., 1995) that has been found in several unrelated patients, including some with *de novo* mutations. This site thus represents a mutational hotspot.

Aldred and Trembath (2000) identified several new mutations in *GNAS1*, among them a case of germ-line mosaicism for a *de novo* in-frame insertion in exon 13. This was one of three insertions and deletions observed at this site, suggesting that nt 1106-1108 in exon 13 represent a second potential "hot spot" for mutations.

A particularly interesting $G_s\alpha$ mutation was described in two patients showing a combination of testotoxicosis (a condition caused by autonomous secretion of testosterone by Leydig cells in the absence of LH, part of the clinical spectrum of McCune-Albright syndrome) and PHP-Ia. In both patients, a $G_s\alpha$ mutation replaced

alanine at position 366 with serine. This mutation constitutively activates adenylyl cyclase *in vitro*, causing hormone-independent cAMP accumulation when expressed in cultured cells, and accounting for the testotoxicosis phenotype (as cAMP stimulates testosterone secretion). Although $G_s\alpha$-A366S is stable at testis temperature, however, it is rapidly degraded at 37°C, explaining the PHP-Ia phenotype caused by loss of $G_s\alpha$ activity. *In vitro* experiments indicated that accelerated release of GDP causes both the constitutive activity and the thermolability of A366S (Iiri et al., 1994).

Although these mutations provided a molecular basis for the $G_s\alpha$ deficiency, they did not offer an explanation for why an identical defect in $G_s\alpha$ should have such variable phenotypic consequences in different tissues or in different individuals (Miric et al. 1993).

3.1.7.2.5 PHP-Ia PATTERN OF INHERITANCE AND GENOMIC IMPRINTING

One of the most interesting and puzzling characteristics of PHP-Ia has been its pattern of inheritance and the relationship of this to the spectrum of clinical characteristics observed among the members of the families affected. The pattern of inheritance was difficult to establish because of an apparent excess of affected females and the presence within an affected family of members with the same somatic features (AHO) but no evidence of hormone resistance (PPHP). The question raised was: "how is possible that the same mutation within one family may have a different phenotypic expression in different affected individuals?".

Levine et al. (1986) proposed that the expression of the mutations depends partly upon modifying genes, invoking the concept of metabolic interference. This proposes that two different non-allelic genes, which by themselves may cause no overt metabolic abnormalities, could interact when present together to cause hormone resistance. Thus, a patient with PHP-Ia would not only have reduced $G_s\alpha$ activity, but might also have inherited non-allelic polymorphisms, resulting in altered activity

of other subunits of e.g. adenylyl cyclase, cAMP phosphodiesterase or protein kinase. Conversely, patients with PPHP might inherit only reduced $G_s\alpha$ activity and thereby escape expression of overt endocrine dysfunction.

Controversially, early clinical studies indicated a female:male ratio of 2:1 among affected individuals, which led to the suggestion that AHO was an X-linked disorder. The subsequent description of a family with male-to-male transmission suggested that PHP-Ia had an autosomal dominant pattern of inheritance; however the features exhibited by the proband and his sister were not striking and difficult to explain. The authors noted that for unknown reasons the pattern of inheritance albeit autosomal dominant had an enhanced expression in the female (Weinberg et al., 1971). Nonetheless, the previous observation and the subsequent location of the human *GNAS1* gene to chromosome 20 led to the conclusion that PHP-Ia must be inherited in an autosomal dominant fashion (Weinberg et al., 1971, Schwindinger and Levine, 1997).

To explain these differences in clinical presentation, it was suggested that they might be related to genomic imprinting, depending on the sex of the transmitting parent (Hall, 1990). In the analysis by Davies and Hughes (1993), a survey of the published records of AHO pedigrees showed that in 36 reports of families with AHO in two or more generations, 33 cases presented maternal transmission and only three were paternal; of 66 affected offspring, 36 were female and 30 were male (in accordance with an autosomal dominant inheritance ratio). The 60 offspring who had inherited the gene from their mother had full expression of the syndrome, with hormone resistance (PHP); the 6 offspring with paternal transmission all had partial expression (AHO only, PPHP). The study suggested that full expression of PHP-Ia was associated with inheritance of a maternally transmitted mutation, while inheritance of a paternally transmitted mutation results in partial expression irrespective of the degree of expression in the parent. The suggestion was made that *GNAS1* might therefore be an imprinted gene, with loss of a maternally-expressed function accounting for hormone resistance in PHP-Ia.

Further support for this hypothesis came from the study by Wilson et al. (1994), they studied the origin of a $G_s\alpha$ mutation in apparently sporadic cases of AHO, demonstrating that the "switch" in the phenotype was dependent on the sex of the transmitting parent across two generations. They remarked the fact that simple monoallelic expression of the *GNAS1* gene alone would not explain the clinical observations in AHO, as equivalent reductions in $G_s\alpha$ bioactivity were found in bothPHP-Ia and PPHP patients in all tissues tested. However, imprinting at a tissue- or even cell-specific level could account for these findings.

One contradictory report came from Schuster et al. (1994), who reported a family with PHP-Ia that had inherited the disease paternally as well as maternally, suggesting that mechanisms other than genomic imprinting are responsible for the full expression of hormone resistance, at least in the case of this family.

At a molecular level, demonstration that *GNAS1* was biallelically-expressed came from the studies by Campbell et al. (1994). They analysed the origin of *GNAS1* transcripts in human fetal tissues, and of 75 fetuses genotyped 13 heterozygous for a *Fok*I polymorphism in exon 5 of *GNAS1* were identified. RNA from up to 10 different tissues from each fetus was analysed by RT-PCR. In all cases expression from both parental alleles was shown by *Fok*I digestion of the resulting fragments. No tissue specific pattern of expression was discerned in these experiments. The authors stated that if genomic imprinting regulates the expression of the human *GNAS1* gene, the data suggested that the effect must either be subtle and quantitative or be confined to a small subset of specialised hormone responsive cells within the target tissues.

As mentioned this paradox could be explained if imprinting alters the expression of $G_s\alpha$ in a cell-specific fashion. Thus the different phenotypes could result from tissue-specific combinations of haploinsufficiency and monoallelic expression. The functional defects found in patients with PHP-Ia reflect $G_s\alpha$ activity in cells in which the gene is subject to paternal imprinting but haplosufficient (one allele normally suffices, for instance in PTH-responsive cells of the kidney). On the other hand, the

phenotype of AHO results from deficient signalling function in cells in which the $G_s\alpha$ gene is haploinsufficient but not imprinted (both alleles are normally expressed). Finally cAMP-dependent functions that are not affected in either PHP-Ia or PPHP are performed by cells in which the $G_s\alpha$ gene is haplosufficient and not imprinted (Farfel et al., 1999).

As alluded to above in the discussion of *GNAS1* gene structure, the imprinting status of *GNAS1* was revaluated following the discovery of a differentially methylated region approximately 35 kb upstream of exon 1. These new studies began with an analysis carried out by a collaborating group in Japan (lead by Dr. Y. Hayashizaki). These investigators used restriction landmark genomic scanning to compare human parthenogenetic DNA, DNA from non-related female individuals and DNA from androgenetic complete hydatidiform mole. A differentially methylated spot, named A20, was identified. When this was cloned and sequenced, analysis of EST databases demonstrated that A20 was part of a cDNA derived from *GNAS1*. The knowledge (i) that imprinted genes carry differences in methylation (ii) that *Gnas* (the homologous gene in mice) maps to a known imprinted region and (iii) the pattern of inheritance observed in families suffering from PHP-Ia, all prompted the study of this new differentially methylated region, in order to define its relationship to *GNAS1* and subsequently to assess its imprinting status.

3.2 MATERIAL AND METHODS

3.2.1 OLIGONUCLEOTIDES

The following oligonucleotides were used in the analysis described in this chapter:

NAME	SEQUENCE 5´- 3´
A20A	ATC TTT GGT GGA TTC GGG TGG
A20B	GGA CAG GCA GGG GCT TCA GGT
A20D	GCC TTC CCC AAT GAC CCC AG
A20F2	GTG TGT GTT GGT GTC CAT ATT C
A21F	GTT TCC TCA CTT CTC ATT C
A21R	CTG TGT GTC TGG TTT TTC AA
A21R2	AGC CCT TCC CAA GTA CAA GTA
Xl1F	GGA TGC CTC CGC TGG TTT CAG
Xl2R	ACT CTC AGA TCG ACC GAA GCA G
XL10	GTC GGG ATC TTC TGG GGT T
XL11	TGA CGC CCC AGC CGA T
K06T7	AGC CGG AAA CGA CCT CAG TG
K06-T3	AGA AAA TGG TGC CGA GCG ACA
K06A	ACT TAC TCC GCA ACT TTC TC
K06B	GTA CCC CTG GCG GAG AAG CG
K06C	CGC AGA GAA GAA ACG CAG T
K06D	ACT GCG TTT CTT CTC TGC G
K06E	CTG AAA CCA GCG GAG GCA TCC
K06F	CTC TCA GAT CGA CCG AAG CAG C
1Fb	CCA TGG GCT GCC TCG GGA ACA
1R	TAG ACC TGC TTG TCC TTC TGC A
$G_s\alpha 2F$	CCA AAG TGT TAA AAT GCC TCC T
$G_s\alpha 2R$	CCC TGC ACA GAT TTG ACA CTT AC
GNASEx3F	AGA GGA CCG CAG GCT GCA A
GNASEx3.5R	CCT CAG ATA ACA GAA TAG TA

GNASEx5.5R2	TAG AAG TTC CTA AGA CCG AGC AAT TCC AGC
PHP7F	TTG ATG TAG CGC TGT GAA CA
PHP7R	TGC CAA TCT AAA GCA AAG GTC
F8Bg1M13F-1	TTT CCT CGG GGG TTG GTA T
F8Bg1M13R-1	TCC CTC CTG CCA CCT GCC
F8F2	TAC ATT GCT TAC AGT TCA GGA A
I01F	AGG CAA GAG GAC CGG CGG AGG
I01R	GCT CCG CGC CTT GCT TGC CTC
NESP1	TCG GTG TGT GCC TAA GAG GAT
NESP2	CTT GGA TCT TTG TCC TCG GG
NESP3	GTC ACT AAT GGA GGA CGC CGT
NESP4	TGG TGG GCG TTA AGG AAG CT
NESP5	TCG GAA TCT GAC CAC GAG CA
NESP6	CGT CTG CAC GCT CTC AAG TT
NESP7	CAG GCT CGG TCT CGA AGT
NESP8	CGG GGG ACT CAG GGA ATA CC
NESP9	CTC CGA TCC ATC CTC TT
NESP10	CCC GAA CGC ACC CCA GAT
NESP11	AGT GAC GCC GGA TGG GGA
NESP12	ATT TGG GGT GCG TTC GGG
NESP13	TTC CCC TAA CCG CCC GCC G
NESP14	CGG CGG GCG GTT AGG GGA A
NESP15	TTC CTG AAC TGT AAG CAA TGT A
NESP16	TTC CTG AAC TGT AAG CAA TGT A
NESP17	GGG GTC GCG TCT AAC ATC A
NESP18	TGA TGT TAG ACG CGA CCC
NESP21	AAA AAT CTC ATC CGA CTT GG
NESP22	CCA AGT CGG ATG AGA TTT TT
NESP23	AGC CTG AGT TCA CCA AGC A
NESP24	TGC TTG GTG AAC TCA GGC T
NESP25	GTC ATT CAT CTC AAG CCC T
NESP26	CGG CTG CTC TCA CAT CGT

3.3 RESULTS

3.3.1 IDENTIFICATION OF EXON A20 OF *GNAS1*

A screen for genomic regions displaying differential allelic methylation (DMRs) was carried out in collaboration with the laboratory of Dr. Y. Hayashizaki (Tsukuba, Japan). This screen used restriction landmark genomic scanning for methylation differences (RLGS-M), a technique developed in the Hayashizaki laboratory (Hatada et al., 1991, Hayashizaki et al., 1994, Kamiya et al., 2000). The RLGS-M method employs ^{32}P-end-labelling of rare-cutting methylation-sensitive restriction sites in genomic DNA, followed by high resolution two-dimensional electrophoresis separation of end fragments generated by cleavage with a second and third restriction enzyme.

This method preferentially targets transcription units because of its dependence on cleavage at rare CpG-containing sites, which are located mostly within CpG islands at the 5′ ends of genes. RLGS-M is dosage sensitive and can distinguish haploid and diploid intensity of single-copy genes. This important property confers the ability to screen for imprinted genes by comparing spot intensity in parthenogenetic, normal and androgenetic DNA, without having to rely on polymorphisms to distinguish the parental alleles. A spot generated by cleavage at a site within a DMR is predicted to be of haploid intensity in all normal DNA samples, but to be of either diploid or zero intensity (or vice-versa) in parthenogenetic and androgenetic DNA, depending on whether the methylated allele is paternal or maternal, respectively.

In this screen, parthenogenetic DNA (from blood of the chimaeric patient FD, Strain et al., 1995), DNA from normal non-related female individuals and from a complete hydatiform mole were compared. Two spots named NV149 and A20 were identified as candidate imprinted loci. These were excised from the gel, and cloned and sequenced. The NV149 clone, which is not discussed further here, proved to map to a region upstream of the *ZAC* gene (the zinc finger protein which regulates apoptosis and cell cycle arrest). This allowed the demonstration that *ZAC* is a paternally

expressed imprinted gene and most likely the gene that is dysregulated in transient neonatal diabetes mellitus (Kamiya et al., 2000).

The other differentially methylated spot, named A20, that was absent from the RLGS-M images from parthenogenetic DNA, but present at haploid intensity in all control females, was also cloned and sequenced. This *Mbo*I-*Asc*I fragment was 232 nt in length. Database searching using this sequence resulted in identification of a match to an EST (Acc. number AJ224867), derived from IMAGE cDNA clone 359933. Alignment of the A20 genomic sequence fragment with the cDNA sequence revealed that the A20 clone contained a single exon that was 91 nt in length (referred to hereafter as exon A20) (Figure 3.3).

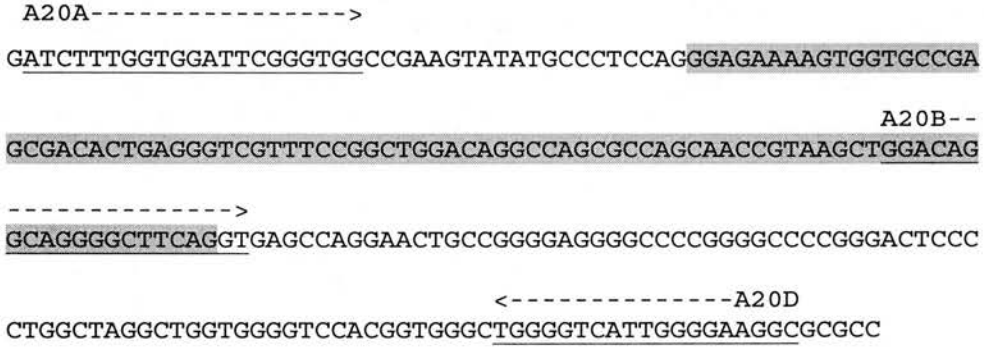


Figure 3.3. DNA sequence of genomic clone A20. The sequence corresponding to the exon A20 (as determined by comparing cDNA and genomic sequences) is shaded. The oligonucleotides used for analysis and their orientations are marked by arrows and the sequences underlined.

3.3.1.1 Analysis of the cDNA clone 359933 (827-k06)

The IMAGE cDNA clone 359933 was obtained from the UK HGMP Resource Centre (under its plate co-ordinates, 827-k06). This cDNA had been cloned in the vector pT7T3D using the enzymes *Not*I and *Eco*RI. Plasmid DNA was therefore purified and digested with these enzymes. The bands obtained showed that the cDNA insert of the clone was ~800 bp in length (Figure 3.4).

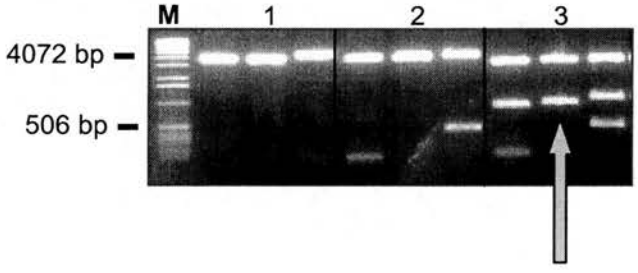


Figure 3.4. Restriction digestion analysis of cDNA clone 359933 (827-k06). Lanes: M) 1 kb marker ladder; 1) *Not*I digest 2) *Eco*RI digest 3) *Not*I+*Eco*RI digest. The arrow marks the insert fragment (the remaining unlabelled lanes correspond to other clones that were being analysed at the same time).

3.2.1.2 Sequence analysis of clone 359933

Clone 359933 was then completely sequenced on both strands; it proved to be 834 nt in length (Figure 3.5). When this complete sequence of the cDNA clone was used to search the sequence databases it was noted that it consisted of:

(i) A 367-nt upstream region, homologous to part of the 5′ portion of the mRNA encoding the rat $G_s\alpha$-related protein, XLαs (Kehlenbach et al., 1994).

(ii) The 91 nt corresponding to exon A20.

(iii) A previously unidentified 67 nt exon (given the name A21).

(iv) *GNAS1* exons 2 and 3.

(v) A variant terminal exon (which was given the name 3N) containing a polyadenylation signal and poly(A) addition site. A similar variant terminal exon has previously been found in other *GNAS1* mRNAs (Crawford et al., 1993).

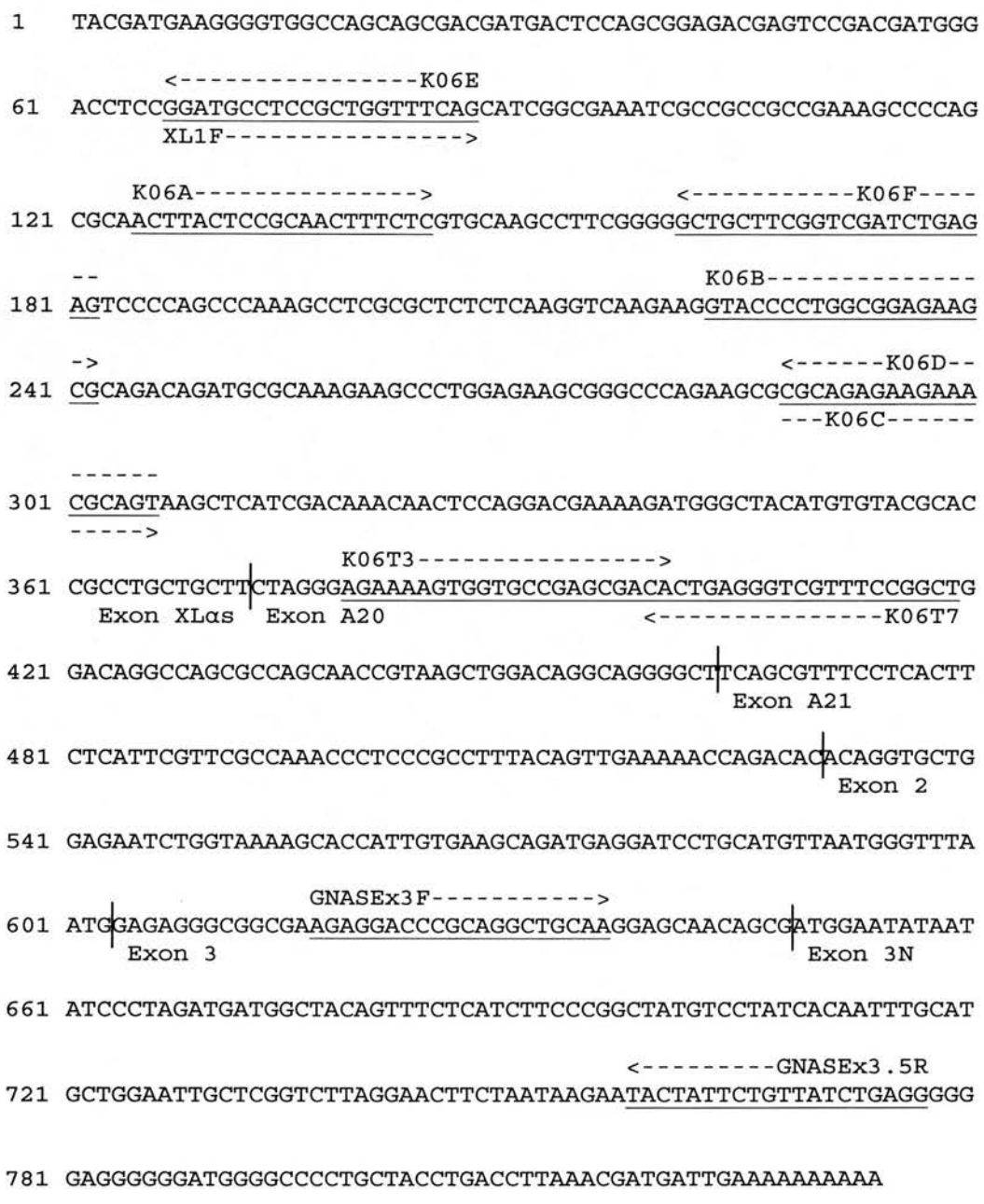


Figure 3.5. Complete sequence of cDNA clone 359933 (827-k06). The splice junction positions are marked by vertical lines. The names of the exons identified are shown. The oligonucleotides used for the analysis and their orientations are marked by arrows and the sequences underlined. This sequence was deposited in GenBank under the accession number AJ224867.

The presence of XLαs-homologous sequences, and of exons 2 and 3 of the *GNAS1* gene, indicated that the sequence of clone 359933 represented a *bona fide GNAS1* transcript, containing at its 5′ end previously undescribed exons, including A20 and A21. The identification of these new exons showed that the previously reported gene structure of *GNAS1* (Kozasa et al., 1988) was incomplete. Furthermore, since A20 had been identified as a candidate imprinted locus, there was clearly a need to characterise the structure of the new 5′ region of *GNAS1*, so as to allow further analysis of the differential methylation in this region and its implications for imprinting of the *GNAS1* locus. As a first step, the RPCI1 PAC DNA library was screened in order to identify large genomic clones that contained *GNAS1*.

3.3.2 SCREENING OF THE RPCI1 PAC LIBRARY FOR CLONES CONTAINING *GNAS1*

For screening the RPCI1 human genomic PAC library by PCR, two oligonucleotides named PHP7F and PHP7R, which match sequences flanking exon 7 of *GNAS1* (as described in Kozasa et al., 1988), were used. Four of the 21 primary pools (named A to U) were positive for the presence of *GNAS1*. These were pools D, E, H and U, all of which yielded a PCR product of the expected size, 180 bp (Figure 3.6A). The secondary pools corresponding to the four positive primary pools were screened using the same PCR primers. Secondary pools D15, E3, H5 and U4 were positive, corresponding to the PAC clone groups numbers 60, 63, 110 and 309, respectively (Figure 3.6B).

Screening of the tertiary pools corresponding to each of these four clone groups identified the following PAC clones as containing exon 7 of *GNAS1*: dJ60o16, dJ63n2, dJ110p14 and dJ309f20 (Figure 3.6C).

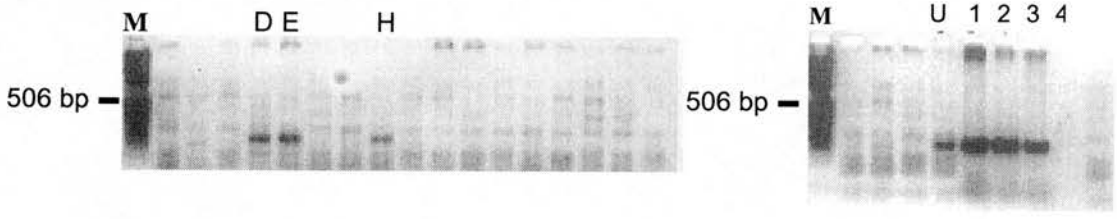
Figure 3.6.

A) PCR screening for *GNAS1* exon 7 in the RPCI1 PAC library. Lanes: M) 1 kb marker ladder; positive pools indicated by the corresponding letter: D, E, H, and U; 1, 2 and 3) human genomic DNA (positive controls); 4) negative control (DNA blank).

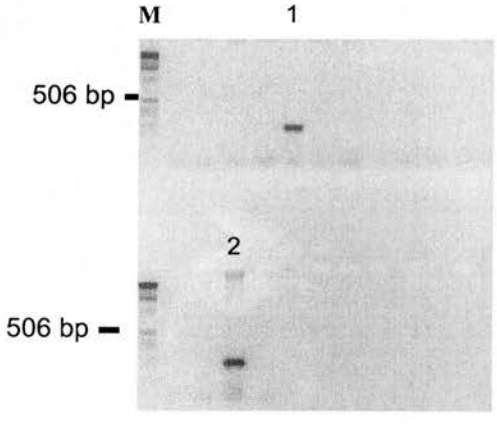
B) PCR screening of secondary pools of the RPCI1 PAC library. The figure shows screening of the group H secondary pools. There was only one positive pool in this group. Lanes: M) 1 kb marker ladder; 1) positive sample (corresponding to clone group number 110); 2) normal human genomic DNA, positive control.

C) Tertiary screening of the RPCI1 PAC library. The PCR products corresponding to PAC clone dJ309f20 are shown. Lanes: M) 1 kb marker ladder; 1) positive sample corresponding to group f; 2) positive sample corresponding to sample number 20; 3 and 4) positive controls (normal human genomic DNA).

A) PCR screening of primary pools of the RPCI1 PAC library



B) Secondary screening, pool H



C) Tertiary screening, pool U4

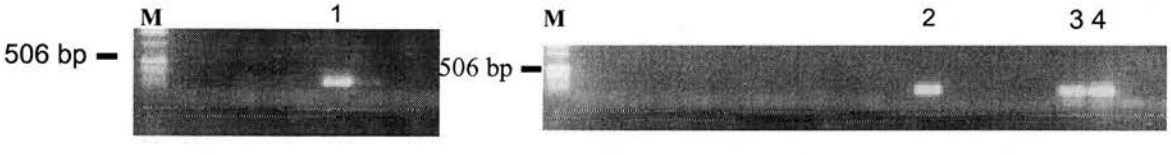
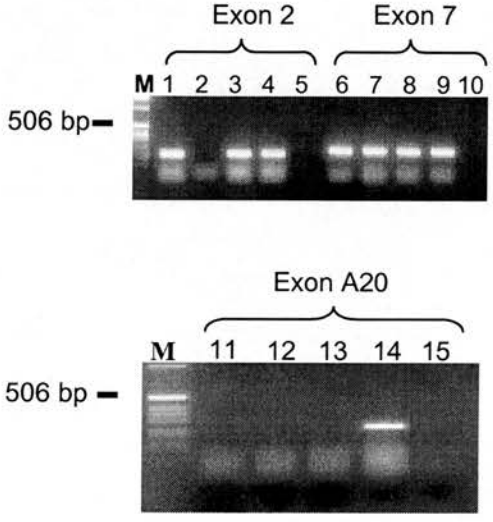


Figure 3.6. PCR screening of the RPCI1 PAC library

These four PAC clones were obtained from the UK HGMP Resource Centre and DNA was extracted. To confirm the presence of *GNAS1* in each clone, two different PCRs were performed. In addition to primers PHP7F and R that flank exon 7, primers Gsα2F and Gsα2R, that flank exon 2 of *GNAS1*, were used. The expected PCR product was 149 nt in length. PCR products of the expected size were obtained from all the PAC clones with one exception; dJ63n2 was negative for the presence of exon 2 (Figure 3.7).

In a third PCR, primers A20A and A20D, that match sequences of the recently identified exon A20 at the 5′ end of *GNAS1*, were used (see Figure 3.3). This time, only dJ309f20 was positive for the expected 222 nt PCR product, establishing this as the only one of the four PACs likely to contain the novel 5′ exons of *GNAS1* (Figure 3.7).



LANE	NAME	EXON 2	LANE	EXON 7	LANE	EXON A20
1	dJ60o16	+	6	+	11	-
2	dJ63n2	-	7	+	12	-
3	dJ110p14	+	8	+	13	-
4	dJ309f20	+	9	+	14	+
5	Control blank		10		15	

Figure 3.7. PCR analysis of the PAC clones for exons 2, 7 and A20 of *GNAS1*. Lane M) 1 kb marker ladder.

3.3.3 STRUCTURAL ANALYSIS OF THE 5′ END OF *GNAS1*

3.3.3.1 Construction of subclone libraries derived from PAC dJ309f20

The first step taken in order to characterise the 5′ end of *GNAS1*, with respect to the newly identified exons, was to create a number of different subclone libraries from PAC dJ309f20. These were then screened to identify subclones containing the new exons. These subclones were then used to establish the genomic structure of *GNAS1* at its 5′ end.

Subclone libraries derived from PAC dJ309f20 were constructed using vectors pCRII (Invitrogen), TVEC (Promega) and pUC18 (Stratagene). PAC dJ309f20 was digested with the enzymes *Not*I+*Sac*I and *Sac*I+*Asc*I for the subclone libraries using the first two vectors. *Bam*HI-digested pUC18 was used to subclone *Bgl*II and *Bam*HI fragments of the PAC.

3.3.3.1.1 Screening for subclones containing the A20 exon

As indicated in Figure 3.3, an *Asc*I site lies just downstream of the A20 exon. From the TVEC vector library, twenty randomly chosen *Sac*I-*Asc*I subclones were screened by colony PCR (using primers A20A and A20D) for the presence of exon A20 of *GNAS1*. Six subclones (numbers 4, 5, 10, 13, 16 and 21) were positive (Figure 3.8).

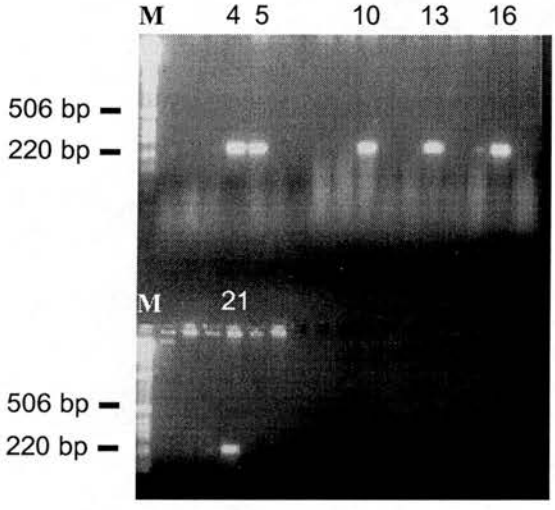


Figure 3.8. Colony PCR screening for exon A20. Lanes: M) 1 kb marker ladder; 4, 5, 10, 13, 16 and 21) subclones positive for the presence of exon A20.

3.3.3.1.2 Screening for subclones containing GNAS1 exon 1

The pCRII subclone library (of *Not*I-*Sac*I fragments) was screened by colony PCR using primers 1Fb and 1R. Two subclones were positive (not shown).

3.3.3.1.3 Screening for subclones containing exon XLαs

Using the *Bam*HI subclone library of dJ309f20, colony PCR was performed using primers XL1F and KO6D, that match sequences in the XLαs part of the 359933 cDNA (Figure 3.5). Of 100 subclones analysed, only two, numbers 21 and 33, yielded the expected 241 nt PCR product. DNA was extracted from both clones and rescreened for both exons A20 (using primers A20A and A20D) and XLαs (primers XL1F and KO6D). Both clones were positive for the presence of A20 in addition to XLαs.

These two subclones (21 and 33) were further analysed using PCR primers A21F and A21R, that match sequences of exon A21. A product of the expected size (67 nt) was again obtained from both subclones. Thus, the XLαs, A20 and A21 exons appeared all to lie on a common *Bam*HI fragment, contained in subclones Bam21 and Bam33.

3.3.3.2 Sequencing analysis of subclone Bam21

Digestion of subclone Bam21 with *Bam*HI showed an insert of ~2 kb. The complete sequence of this insert was then determined on both strands. As expected from the previously described PCR analyses, this subclone contains the complete sequences of exons A20 and A21 and also the 3′ part of the XLαs exon. In order to complete the sequence of the entire region comprising these three exons, further sequencing of *Sac*I-*Asc*I subclone 4, containing exon A20 and exon XLαs, was also carried out in our laboratory (Dr. B. E. Hayward). These overlapping subclones yielded a combined sequence of 2479 nt, which was deposited under accession number AJ224868 (Figure 3.9).

Figure 3.9. Sequence assembled from subclone Bam21 and *Sac*I-*Asc*I subclone 4 (accession number AJ224868). The exons A20 and A21 are shaded. The intron sequences are shown in lower case. The translated part of the XLαs exon probably starts at nt 318 and is shaded (see below). Some of the oligonucleotides used for the sequence analysis and their orientations are shown. Restriction enzymes sites are in bold and underlined.

```
1     GAGCTCCTCCCGAGGAGCCCCAAGCCCTCAGGCCTGCAAAGGCTGGCTCCAGAGGAGGCT
       SacI
61    ACAGCCCTCCCCCTGAGGAGACTATGCCATTTGAGCTTGATGGAGAAGGATTTGGGGACG
121   ACAGCCCACCCCCGGGGCTTTCCCGAGTTATCGCACAAGTCGACGGCAGCAGCCAGTTCG
181   CGGCAGTCGCGGCCTCGAGTGCGGTCCGCCTCACTCCCGCCGCGAACGCGCCTCCCCTCT
241   GGGTCCCAGGCGCCATCGGCAGCCCATCCCAAGAGGCTGTCAGACCTCCTTCTAACTTCA
                      XLαs
301   CGGGCAGCAGCCCCTGGATGGAGATCTCCGGACCCCCGTTCGAGATTGGCAGCGCCCCCG
361   CTGGGGTCGACGACACTCCCGTCAACATGGACAGCCCCCCAATCGCGCTTGACGGCCCGC
421   CCATCAAGGTCTCCGGAGCCCCAGATAAGAGAGAGCGAGCAGAGAGACCCCCAGTTGAGG
481   AGGAAGCAGCAGAGATGGAAGGAGCCGCTGATGCCGCGGAGGGAGGAAAAGTACCCTCTC
541   CGGGGTACGGATCCCCTGCCGCCGGGGCAGCCTCAGCGGATACCGCTGCCAGGGCAGCCC
            BamHI
                                          <--------------XL10
601   CTGCAGCCCCAGCCGATCCTGACTCCGGGGCAACCCCAGAAGATCCCGACTCCGGGACAG

661   CACCAGCCGATCCTGACTCCGGGGCATTCGCAGCCGATCCCGACTCCGGGGCAGCCCCTG
721   CCGCCCCAGCCGATCCCGACTCCGGGGCGGCCCCTGACGCCCCAGCCGATCCCGACTCCG

                 XL11---------->
781   GGGCGGCCCCTGACGCCCCAGCCGATCCAGATGCCGGGGCGGCCCCTGAGGCTCCCGCCG

841   CCCCTGCGGCTGCTGAGACCCGGGCAGCCCATGTCGCCCCAGCTGCGCCAGACGCAGGGG
901   CTCCCACTGCCCCAGCCGCTTCTGCCACCCGGGCAGCCCAAGTCCGCCGGGCGGCCTCTG
961   CAGCCCCTGCCTCCGGGGCCAGACGCAAGATCCATCTCAGACCCCCCAGCCCCGAGATCC
1021  AGGCTGCCGATCCGCCTACTCCGCGGCCTACTCGCGCGTCTGCCTGGCGGGGCAAGTCCG
1081  AGAGCAGCCGCGGCCGCCGCGTGTACTACGATGAAGGGGTGGCCAGCAGCGACGATGACT
                 NotI
                                       XL1F---------------->
1141  CCAGCGGAGACGAGTCCGACGATGGGACCTCCGGATGCCTCCGCTGGTTTCAGCATCGGC
                                       <----------------K06E

1201  GAAATCGCCGCCGCCGAAAGCCCCAGCGCAACTTACTCCGCAACTTTCTCGTGCAAGCCT

            <------------------XL2R
1261  TCGGGGGCTGCTTCGGTCGATCTGAGAGTCCCCAGCCCAAAGCCTCGCGCTCTCTCAAGG

1321  TCAAGAAGGTACCCCTGGCGGAGAAGCGCAGACAGATGCGCAAAGAAGCCCTGGAGAAGC

                KO6C----------------->
1381  GGGCCCAGAAGCGCGCAGAGAAGAAACGCAGTAAGCTCATCGACAAACAACTCCAGGACG
                <BssHII-----------K06D

1441  AAAAGATGGGCTACATGTGTACGCACCGCCTGCTGCTTCTAGgtaatgcggcggactctg
1501  cctgcgggcagcagggccgccggggaaccggggaggggtggcagggctgcctggtgggg
1561  ctaggggctccgcagtgggaggaggggtccagccaaaggcgggaagaacttgctagaaa

                                                  A20A---------------
1621  gttccagggagggaccctaactgccctgggagcgggaggggatctttggtggattcgggt
      ->                      A20
1681  ggccgaagtatatgccctccagGGAGAAAAGTGGTGCCGAGCGACACTGAGGGTCGTTTC
1741  CGGCTGGACAGGCCAGCGCCAGCAACCGTAAGCTGGACAGGCAGGGGCTTCAGgtgagcc

1801  aggaactgccggggagggccccgggccccgggactcccctggctaggctggtggggtc
            <---------------A20D
1861  cacggtgggctggggtcattggggaaggcgcgccccgcctcgcctggcacggctgcttgg
                                  AscI

                          A20F2---------------->
1921  actcagacagcttgtcgttggtgtgtgttggtgtccatattctgtgttcgcctgtgcatg

                                                                  <-
1981  atacgctgaagtgtcccctgttcctatcccggcacccccaaattacccgccgactgtgta

      ---------------A21R2
```

```
2041  cttgtacttgggaacgggctgtgttgtggtttgcgccatggcccgggcaaaaaactttcg

2101  ctgttcgcacactctggtggtacctgcgcccgggcgctctgcggcaagcggtgcggagga
2161  cacgcggggaaggtggcggggcctcccgggaaataagcggggcccttggcctgggggagc

2221  ggggaatcgcttttcgccggcctccgcgtaaccttgtttctgtatgtctttctctctttt
                      A21F---------------->
2281  tccttttgcgctaagCGTTTCCTCACTTCTCATTCGTTCGCCAAACCCTCCCGCCTTTAC

2341  AGTTGAAAAACCAGACACACAGgtatctggggcgcccggggctgcacacctcagttcggt
        <---------------A21R
2401  attctgcacacacacagctcgcctggtcgggggatgggggaaaggggaaacacaacctgc

2461  tgtaggcaagtccggatcc
                  BamHI
```

The true 5′ end of exon XLαs remains to be determined, as attempts to map the 5′ end of XLαs by RACE were not successful (Dr. B. E. Hayward). However, RT-PCR products diagnostic of XLαs-exon 2 spliced mRNA could be generated by using upstream primers as much as 1.1 kb from the 3′ end of the XLαs exon (nt 375-1482 in Figure 3.9). These experiments indicate that the XLαs exon is > 1.1 kb in size, and it may extend even further 5′ because the reading frame remains open beyond the start of our sequence (Figure 3.9).

3.3.3.3 Identification of exon 3N of *GNAS1*

As mentioned above, an alternative terminal exon sequence was identified in IMAGE clone 359933 (827-k06). This exon was located after exon 3, and was given the name 3N. To establish the genomic distance between exon 3 and exon 3N, primers GNAS3ExF (exon 3) and GNAS3.5R (exon 3N) were designed and used for PCR analysis. Another oligonucleotide 3.5FII was also used to screen the *Bam*HI and *Bgl*II subclone libraries for subclones containing exon 3N, yielding one positive *Bam*HI subclone (F8).

It was subsequently noticed that primer 3.5FII had been synthesized with a mismatch in its sequence. Nonetheless, when a new primer with the correct sequence was used in a PCR reaction for exon 3N, subclone BamF8 was indeed positive for the expected product of 113 bp. To confirm the presence and sequence of exon 3N, subclone BamF8 DNA was then used as template for sequencing with primers GNASEx3.5, GnasEx3.5R and 3.5FII.

These data were used to determine the position and size of exon 3N. The PCR product obtained using primers in exons 3 and 3N was ~1100 bp (Figure 3.10). Thus, exon 3N was determined to be 1 kb downstream of exon 3, within the ~4.5 kb intron originally described between exons 3 and 4 (Kozasa et al., 1988). There were no additional splice junctions within this region of the IMAGE clone downstream of exon 3.

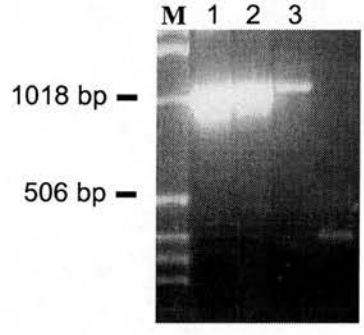


Figure 3.10. GNAS3F-3.5R PCR product. Lanes: M) 1 kb marker ladder, 1) Subclone BamF8 , 2) PAC dJ309f20; 3) human genomic DNA (positive control).

3.3.3.4 Organisation of novel *GNAS1* exons

Based on the previously described results, it was established that the organisation of the novel exons contained within IMAGE clone 359933 (827-k06) is as follows (Figure 3.11):

Figure 3.11. Organisation of the novel exons of *GNAS1*. The exons are in grey or white for untranslated regions.

A) *GNAS1* partial gene structure The exon labeled A has been previously described (Swaroop et al., 1991), the exon 3N is indicated. *F* marks the polymorphism used for analysing the RT-PCR products (described in section 3.3.5).

B) The gene structure of *GNAS1* including the XLαs exon ~35 kb upstream of the originally described exon 1 of *GNAS1* is shown.

C) Organisation of the XLαs region. The XLαs, A20 and A21 exons are shown. The restriction sites indicated are S, *Sac*I; B, *Bam*HI; As, *Asc*I. The region sequenced in subclones Bam21 and *Sac*I-*Asc*I4 is indicated (modified from Hayward et al., 1998a).

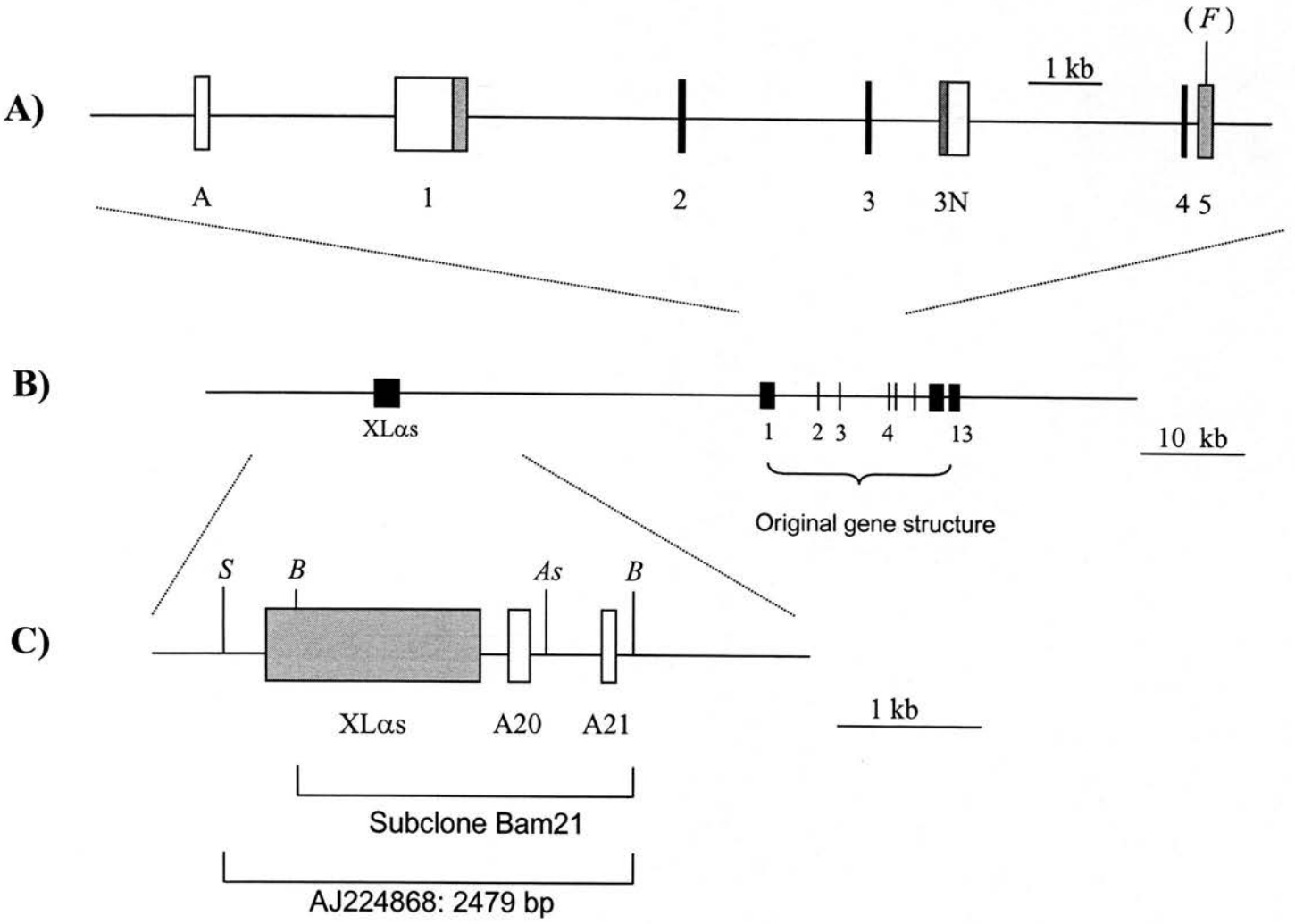


Figure 3.11. Organisation of the novel exons of *GNAS1*

The distance between the XLαs region and the exon 1 region of the gene was established to be approximately 35 kb by restriction mapping of PAC dJ309f20, performed by pulsed-field gel electrophoresis by Dr. B. E. Hayward. Subsequent sequence data of the entire region, obtained at the Sanger Centre, have confirmed these analyses; *GNAS1* exons 1-13, XLαs and *NESP55* (section 3.3.6) lie within the sequence of PAC clones dJ543j19 (Acc. number AL109840), dJ309f20 (Acc. number AL121917), and dJ806m20 (Acc. number AL132655).

3.3.4 DIFFERENTIAL METHYLATION IN THE XLαs REGION OF *GNAS1*

The genomic sequence obtained in this study shows that the human XLαs-homologous region is encoded in a single large exon (>1.1 kb) ~35kb upstream of exon 1 of *GNAS1*. A 1.9 kb *Sac*I-*Asc*I probe fragment (see Figure 3.11) was used for assessing the methylation status of the *Not*I, *Bss*HII and *Asc*I sites in this region, by Southern blotting of normal and parthenogenetic lymphoblastoid cell DNA (Dr. L. Strain). *Sac*I digestion generates a fragment of ~6 kb containing the XLαs exon.

The results of this experiment showed that normal DNA yielded bands indicating the presence of both methylated and unmethylated alleles at all the mentioned restriction sites. The parthenogenetic DNA, in contrast, was completely methylated at all the sites. This confirmed that as indicated by the original RLGS-M result, this genomic region is indeed differentially methylated, the paternal allele being unmethylated and the maternal methylated. The restriction sites whose methylation analysis was assessed are shown in Figure 3.12.

Figure 3.12. Organisation of the XLαs region of *GNAS1*. The map is not to scale. Grey lines above the exons represent the splicing patterns observed in the majority of XLαs-containing RT-PCR products and below the line the splicing patterns seen in the 359933 cDNA clone. Italic letters indicate restriction sites: *S*, *Sac*I; *B*, *Bam*HI, *N*, *Not*I; *As*, *Asc*I; *Ng*, *Ngo*MI; *Bs*, *Bss*HII; *F*, *Fok*I (polymorphic site). Asterisks indicate those sites whose methylation status was assessed by Southern blotting. The coding regions of exons are shaded grey, and the untranslated regions white.

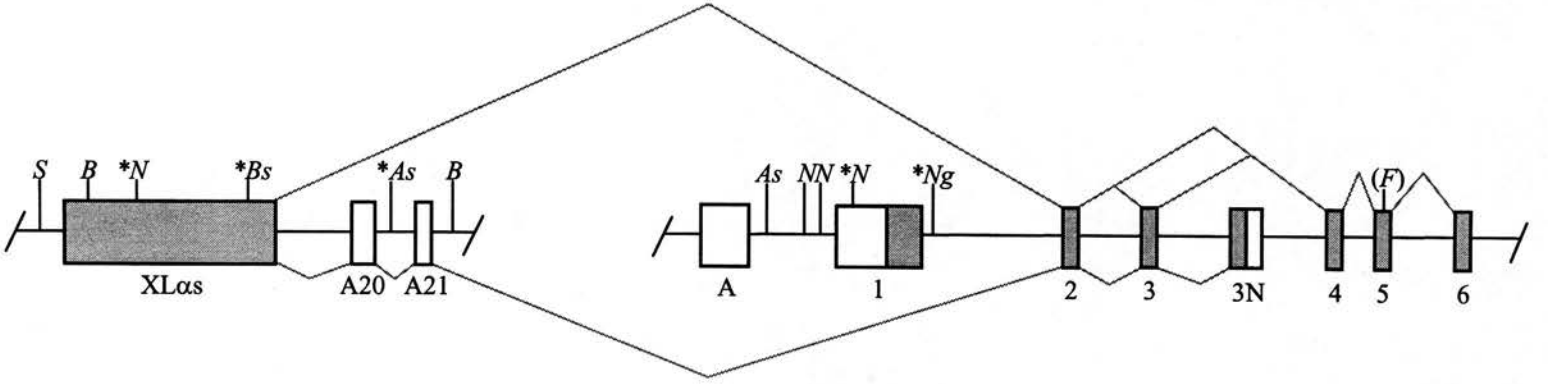


Figure 3.12. Organisation of the XLαs region of *GNAS1* and locations of the restriction sites analysed for differential methylation

3.3.5 MONOALLELIC EXPRESSION OF XLαs-CONTAINING TRANSCRIPTS

In order to assess if the differential methylation described in *GNAS1* reflects monoallelic expression of this gene, a series of RT-PCR analyses of fetal RNA samples was carried out in our laboratory (Dr. B.E. Hayward). In a previous analysis of *GNAS1* described by Campbell et al. (1994) a large number of RNA samples from first trimester fetuses heterozygous for a *Fok*I polymorphism in exon 5 had been analysed. RT-PCR analysis using oligonucleotides in exons 2 and 6 of *GNAS1* had demonstrated that this gene was biallelically expressed in all the samples. However, it was now recognised that these experiments did not distinguish between *GNAS1* transcripts containing either exon 1 or the novel differentially methylated XLαs exon.

In the new study, these fetal RNA samples were first reanalysed by RT-PCR between exon 2 and 6, yielding identical results as previously described, with uniformly biallelic expression of the gene. However when the RT-PCR analysis was performed using an upstream oligonucleotide within the XLαs exon, the *Fok*I digests revealed that all of the XLαs-containing transcripts were derived from the paternal allele. In contrast with this result, an RT-PCR specific for exon 1-derived products showed that these transcripts were derived from both alleles.

These RT-PCR experiments also showed that the XLαs-exon 6 RT-PCR products were shorter than the 783 bp product predicted from the sequence of the cDNA clone 359933. When these products were sequenced, it was found that as in the rat XLαs mRNA, the human XLαs exon is spliced directly onto exon 2, resulting in a mRNA with an open reading frame continuous with that of exons 2-13.

Interestingly, RT-PCR products containing A20 or A21 were not detected, suggesting that most XLαs-containing transcripts do not contain the A20/A21 exons. As A20/A21 transcripts also do not have a continuous ORF, they may be aberrant or functionally unimportant splice forms. However, their RT-PCR analysis (using an

upstream primer specific for A20) did show that A20-containing transcripts also are derived exclusively from the paternal allele.

3.3.6 SEQUENCE CONSERVATION AND REPETITIVE ELEMENTS IN THE XLαs REGION

A comparison between the translated human XLαs ORF (starting at nt 318 of our genomic sequence AJ224868) and the peptide sequence predicted by a full-length rat cDNA (according to Kehlenbach et al., 1994, and in Hayward et al., 1998a) demonstrates that the extreme N- and C-terminal parts of the XLαs segments are well conserved. The central regions, in contrast, align much less well, but there are qualitative similarities between the two species' sequences in this region.

As shown in Figure 3.13, this central region includes a repetitive domain. In the rat XLαs protein, there is a tetrapeptide repeat based on EPAA (single-letter amino acid code). The predicted human protein has a tripeptide repeat motif, again rich in proline, alanine and an acidic amino acid (though D, aspartate rather than E, glutamate). A higher order repeat of 9 or 12 amino acids can also be discerned within this tripeptide repeat region. The overall ratio of (D+E):P:A in this region (20:22:42) is virtually identical to that in the rat EPAA repeat region.

The fact that the qualitative properties of the repeated XLαs domain, but not its primary structure, are conserved between human and rat, may be significant. It may indicate selective pressure exerted at the level of the protein, but for the general chemical nature of the domain rather than its precise sequence. Alternatively, it could be that the presence of the DNA repeat motif itself, rather than the encoded protein sequence, is selected for. This could be the case if the repeat motif has a specific role in the imprinting process, as has been suggested from the finding of such tandem repeats in a number of imprinted genes (Neumann et al., 1995a, b).

Figure 3.13. Internal repeat motifs in the human XLαs exon, residues 74 to 225 (as in Hayward et al., 1998a). The most highly conserved residue at each position of the repeat is shaded.

Summarising the information obtained up until this point in the analysis of *GNAS1*, it was concluded that this gene was indeed imprinted, and that this imprinting was promoter-specific. The various promoter-specific transcripts of *GNAS1* are alternatively spliced onto the splice acceptor site of exon 2. In the case of the XLαs-derived transcripts, this yields mRNA species that are derived exclusively from the paternal allele, including one that encodes a protein homologous to the large $G_s\alpha$-related protein XLαs of the rat. The XLαs promoter is active only on the paternal allele, even in the same tissue samples in which transcription from the major promoter (upstream of exon 1) is biallelic.

3.3.7 IDENTIFICATION AND ANALYSIS OF NESP55

The demonstration that *GNAS1* was indeed imprinted, in a promoter-specific fashion, was of considerable interest, but still did not offer an explanation for the imprinted inheritance pattern of hormone resistance in PHP-Ia. The fact that the mutation must be maternally derived for hormone resistance to develop appeared to indicate that *GNAS1* would be maternally expressed, at least in some tissues. The imprinted XLαs transcript, in contrast, is paternally expressed. Indeed, one would predict that patients with PPHP as a result of a paternally inherited null mutation of *GNAS1* might be lacking XLαs function in addition to their 50% loss of $G_s\alpha$ activity. One possibility that was suggested by the identification of the novel XLαs promoter 35 kb upstream of the exon 1 promoter was that there might be additional upstream promoters, that could perhaps be maternally expressed. This prompted a search for *GNAS1* mRNAs with 5′ ends that were different from those already known.

Initially, Dr. B. E. Hayward in our laboratory examined all identifiable *GNAS1* EST sequences in the databases. Only one of them (derived from IMAGE cDNA clone 301896) had a 5′ end that had not previously been seen. Secondly, a 5′-RACE analysis was performed on human fetal muscle RNA, using a primer located in *GNAS1* exon 5. Products were size-selected to be larger than expected for exon 1-containing mRNAs, and then cloned.

This library of cloned RACE products was then screened for the presence or absence of the XLαs exon. One XLαs-negative RACE clone was > 1.3 kb in size and was characterised further. After sequencing, its 5′ end proved to include the entire unknown 5′ sequence of IMAGE clone 301896. However, unlike clone 301896, the RACE clone also contained an additional 0.9 kb of sequence inserted between the unknown 5′ sequence and exon 2. The analysis of this region demonstrated that it encoded the human homologue of the bovine NESP55 protein.

Further study of the new NESP55 sequence by PCR assays demonstrated that the entire cDNA of NESP55 was encoded by a single exon, contained within the same PAC clone dJ309f20, that contains the XLαs exon and exon 1-13 of *GNAS1*. After hybridisation of NESP55 and XLαs-specific oligonucleotide probes to pulse-field gel electrophoresis-separated *Mlu*I and *Not*I digests, it appeared that the NESP55 and XLα exons were separated by a small (11 kb) *Mlu*I fragment.

Bovine NESP55 is encoded by mRNAs in which exons 2-13 of *GNAS1* appear to make up the 3′ untranslated region. Just as for bovine NESP55, the stop codon of the human NESP55 ORF is 5′ to the splice junction with exon 2, so that none of the $G_s\alpha$ coding region is included in the NESP55 ORF (Figure 3.14).

Figure 3.14. Schematic representation of NESP55 cDNA clones. The 5′ and 3′ untranslated regions are indicated by lines (the sequence in the 3′ untranslated mRNA corresponding to $G_s\alpha$ is represented by a broken line). G2-G7 are the sequences corresponding to exons 2-7 of $G_s\alpha$. The NESP55 open reading frame (ORF) is boxed. The four different clones (a, b, c, d) represented were described by Ischia et al. (1997), and derived from bovine adrenal medulla. Their sequences are identical in the 5′ untranslated region and the ORF but differ in the 3′ untranslated region. The positions of the components of a homologous human EST clone (301896) are given below in relation to the NESP55 sequence (modified from Ischia et al., 1997).

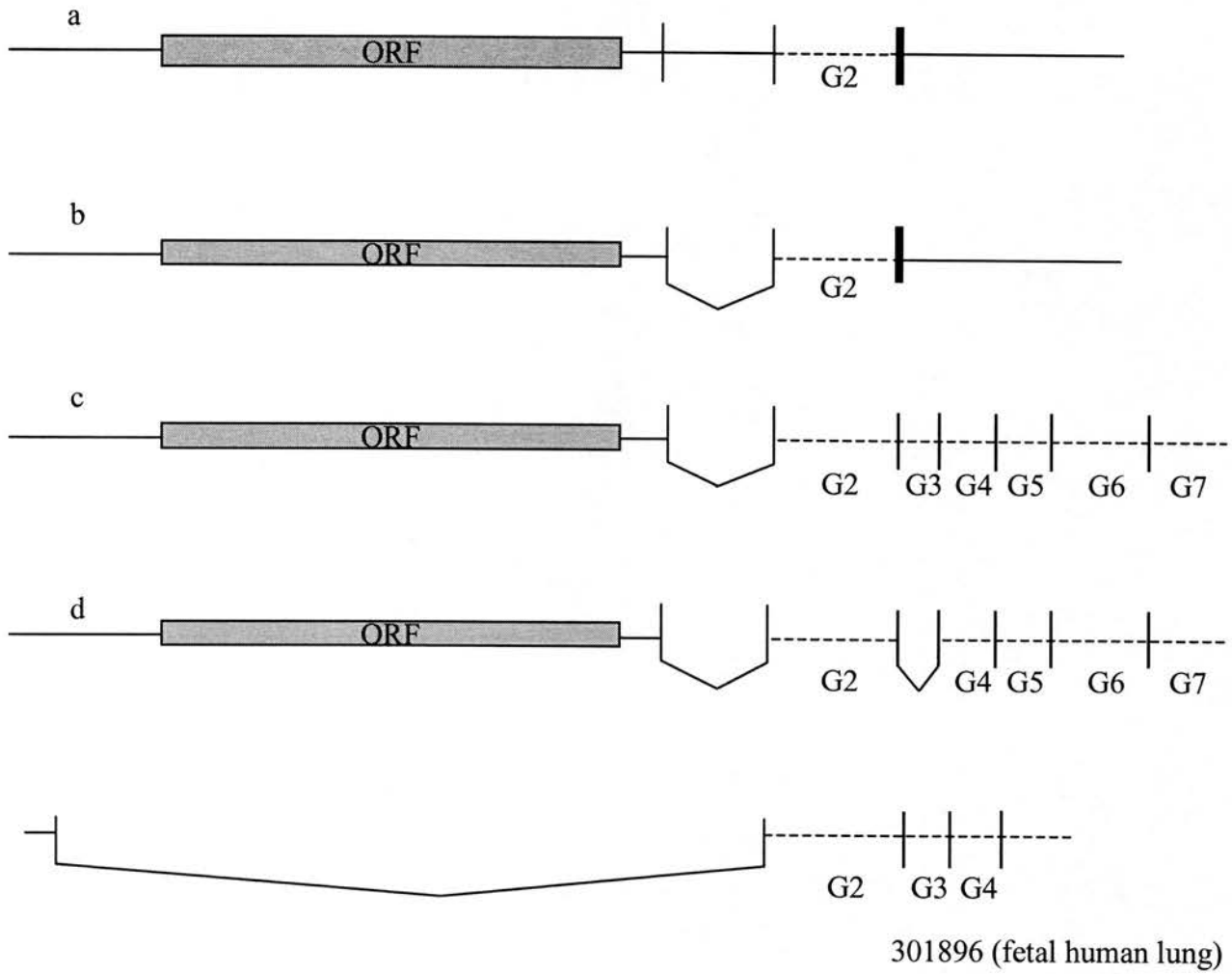


Figure 3.14. Schematic representation of NESP55 cDNA clones

Next, the dJ309f20 subclone libraries of *Bam*HI and *Bgl*II fragments were screened to identify subclones containing the NESP55 exon. One *Bgl*II subclone, 34 F8 (referred to hereafter as subclone Bgl34) was further characterised by restriction mapping and then by direct sequencing of the entire *Bgl*II-derived fragment.

3.3.7.1 Restriction mapping of subclone Bgl34

A restriction map of subclone Bgl34 was first constructed. Some representative digests are shown in Figure 3.15.

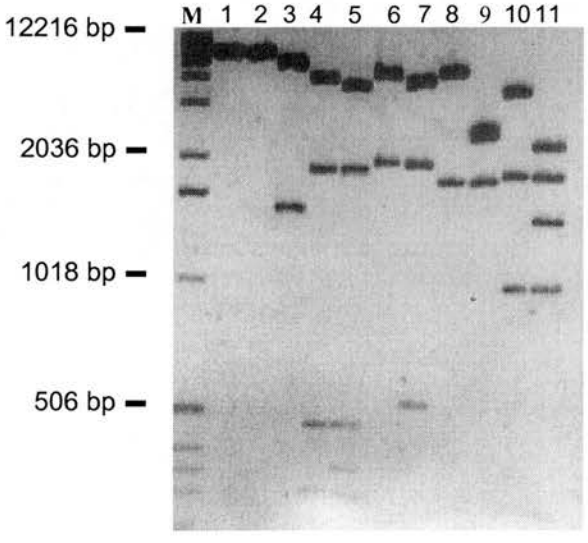


Figure 3.15. Restriction digests of subclone Bgl34. Lanes: M) 1 kb marker ladder; enzymes used for digestion 1) *Kpn*I, 2) *Mlu*I, 3) *Mlu*I+*Kpn*I, 4) *Ngo*MI, 5) *Ngo*MI+*Kpn*I, 6) *Xho*I, 7) *Xho*I+*Kpn*I, 8) *Hin*dIII, 9) *Hin*dIII+*Kpn*I, 10) *Fsp*I, 11) *Fsp*I+*Kpn*I.

The restriction map of subclone Bgl34, deduced from these and other digests, is shown in Figure 3.16.

Figure 3.16. Restriction map of subclone Bgl34 (to scale).

(a) Map of the original PAC clone dJ309f20. (*Left*) The SP6 end of the PAC insert. (*Right*) Exons 1-13 of *GNAS1* (modified from Hayward et al., 1998b).

(b) Expanded view of the Bgl34 insert; *Bgl*II (B), *Mlu*I (M), *Ngo*MI (Ng), *Pvu*I (P), *Sac*II (S), *Fsp*I (F) and *Xho*I (X) sites are indicated.

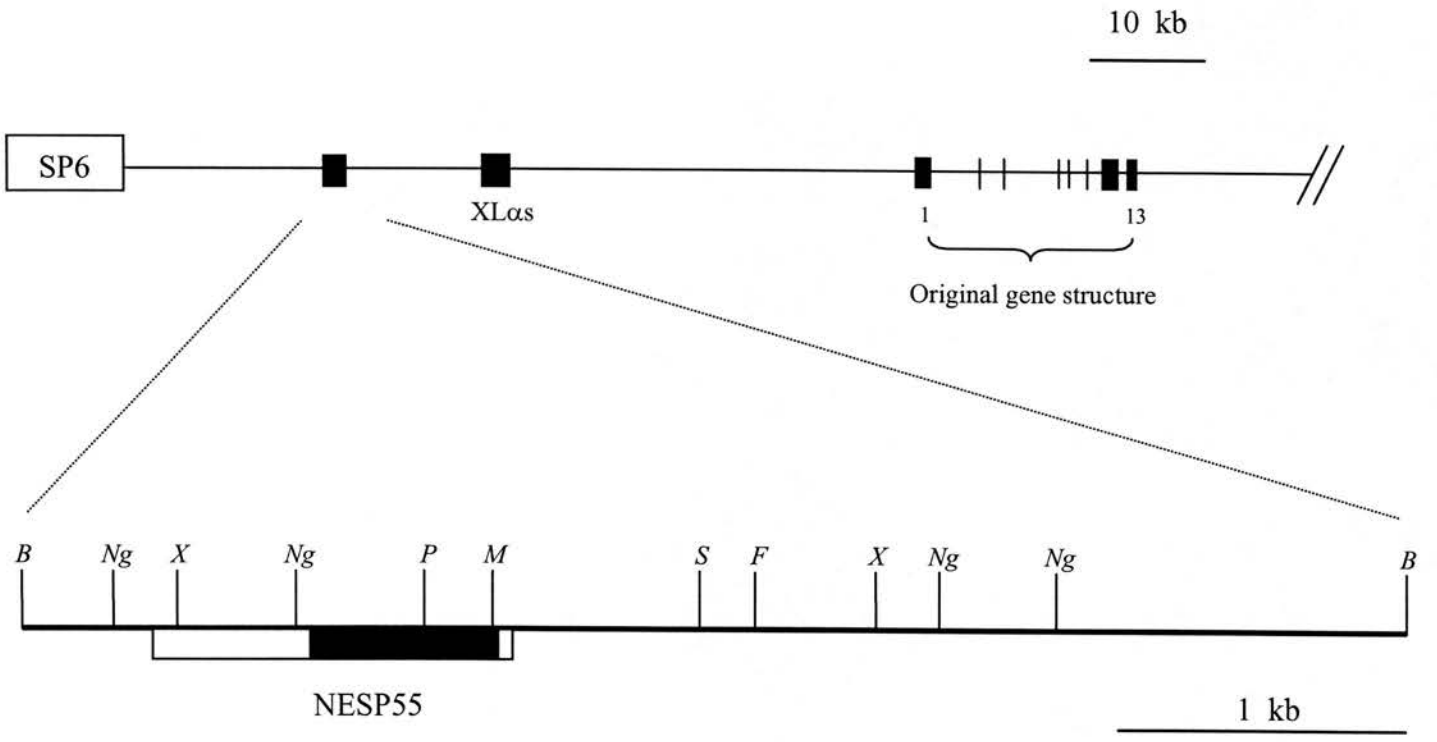


Figure 3.16. Enzymatic restriction map of subclone Bgl34

3.3.7.2 Sequence analysis of subclone Bgl34

The Bgl34 subclone was eventually completely sequenced on both strands. Its insert is 4330 nt in length. The annotated sequence of this subclone is shown in Figure 3.17.

Within this genomic sequence there are at least three regions of interest: ***(i)*** the region containing the upstream sequence common to IMAGE clone 301896 and the 5′-RACE clone previously mentioned (section 3.3.6); ***(ii)*** the NESP55 sequence itself and finally, ***(iii)*** a sequence downstream of NESP55, later identified as corresponding to exon IV of the *GNAS1* antisense transcript (see section 3.3.8 and Chapter 4).

The first of these regions, ***(i)***, spliced in front of *GNAS1* exon 2 in the IMAGE clone 301896 (accession no. W16881) is underlined in Figure 3.17 (nt 567-682). At the end of this region, the presence of a cryptic splice donor site appears to account for the splicing to exon 2. This clone had also been identified by Ischia et al. (1997) when studying the NESP55 bovine sequence (Figure 3.14).

Region ***(ii)*** (the NESP55 ORF) is shown in bold in Figure 3.17. As described above (section 3.3.6), this ORF is contained in a 5′ RACE clone, that also contained upstream sequences, including region ***(i)*** from IMAGE clone 301896. The RACE clone, however, also contained an additional 0.9 kb of sequence, inserted between region ***(i)*** and *GNAS1* exon 2. As is the case also for bovine NESP55, the human NESP55 stop codon is 5′ to the splice junction with exon 2, so that none of the $G_s\alpha$ coding region is included in the NESP55 open reading frame. The putative start codon at nt 849 in Figure 3.17 is the first (and only) one following an upstream in-frame stop codon (nt 685, double-underlined).

The characteristics of the human NESP55 protein sequence are further described in the following section. Further analysis of the genomic sequence of subclone Bgl34 demonstrated that the NESP55 exon is contained within a CpG island that extends

from nt 822 to nt 2839, basically delimited by two *Ngo*MI sites (nt 820-826 and nt 2775-2780). Several restriction sites corresponding to methylation-sensitive rare-cutting enzymes were identified within this CpG island.

Region ***(iii)*** in the Bgl34 genomic sequence is shaded in Figure 3.17. It represents exon IV of an antisense transcript which is further described in section 3.3.8 and chapter 4. This transcript is comprised of 5 exons (numbered I-V) and it has been demonstrated that is imprinted, being expressed from the paternal chromosome only (Hayward and Bonthron, 2000).

Figure 3.17. Complete sequence of subclone Bgl34. The sequence was derived from genomic DNA and initiates at a *Bgl*II site. It is contiguous with the RT-PCR-derived cDNA sequence as far as the splice donor site (indicated by the change to lowercase letters at position 1629), 42 nt beyond the NESP55 stop codon. The NESP55 coding sequence is indicated in bold uppercase letters; the upstream in-frame stop codon is double-underlined. The positions of the ends of 5′ RACE clones, at nt 465, 490 and 517, are indicated by bold underlined nucleotides. The region spliced in front of *GNAS1* exon 2 in IMAGE clone 301896 (Acc. number W16881) is underlined (nt 567-686). The sequence of the cDNA 088f21 corresponding to the antisense transcript exon IV (AS-EXON IV) is shaded, nt 2678-2870 (described in section 3.3.8- and in Chapter 4), and the alternative splice site is indicated by a dashed line at nt 2957 (as described by Hayward and Bonthron, 2000). Some of the restriction sites present in this subclone are marked in bold and italic letters.

```
1       AGATCTGCTAAAGGTAGTTTCATGATTAATGGCTGCCCGAAGTTACCATCCTCAAACTATTAAGCCAT
        BglII

69      TGGAAAATGCTAGAGGCTTAATACTCGTGTATAGACAGACCCTCATTCCCCTGAATATTTGCTATTTG
137     GGGTTCTGGGTTTTTAGCTGCTCCCAGGCTGTGTTTATTTCCCCAATAACTGGGGAAACTGGGGAGGG
205     GCACGACCCCACGGGAGCCTGCGCCCACCTGCCCAAGTACTGGACCTGGGGCTAGCTTGCCGCTTGCT

                                                                F8BglM13R1
273     CCTTGCCACCCGCCAGGCCCCCCGCCCATCGCTTCGCCCAAATCCTTTCAAACAAGGTTCCCTCCTGC

        ------->
341     CACCTGCCGGCCCACTAGGGTCTGCGCCACAGGCTCGGCGCCACCACGCAGCTCGCGGGGAGGTGGCC
             NgoMI
                        <-----------101rII
409     CCCACCTCCTTACTGCACATGCCCGGCAGAAGTCCGGGCGCGCAACTTCGCAGAACCTCACTGCCCGT

                                            I01F---------------->
477     CCCTCCTCGCCTCAGTCTCCTCTGTCCTCTCCCAGGCAAGAGGACCGGCGGAGGCACCTCTCTCGAGT
                                                                  XhoI

545     CTTAGGCTGCGGAATCTAAGACTCAGCGAGAGGAGCCCGGGAGGAGACAGAACTTTCCCCTTTTTTCC

                          <--------------I01R1
613     CATCCCTTCTTCTTGCTCAGAGAGGCAAGCAAGGCGCGGAGCTTTAGAAAGTTCTTAAGTGGTCAGGA

681     AGGTAGGTGCTTCCCTTTTTCTCCTCACAAGGAGGTGAGGCTGGGACCTCCGGGCCAGCTTCTCACCT

749     CATAGGGTGTACCTTTCCCGGCTCCAGCAGCCAATGTGCTTCGGAGCCACTCTCTGCAGAGCCAGAGG

                  NESP1---------------->
817     GCAGGCCGGCTTCTCGGTGTGTGCCTAAGAGGATGGATCGGAGGTCCCGGGCTCAGCAGTGGCGCCGA
            NgoMI                   <----------------
                                            NESP9

885     GCTCGCCATAATTACAACGACCTGTGCCCGCCCATAGGCCGCCGGGCAGCCACCGCGCTCCTCTGGCT
953     CTCCTGCTCCATCGCGCTCCTCCGCGCCCTTGCCACCTCCAACGCCCGTGCCCAGCAGCGCGCGGCTG

                        <--------------NESP4                <------NESP 8------
1021    CCCAACAGCGCCGGAGCTTCCTTAACGCCCACCACCGCTCCGGCGCCCAGGTATTCCCTGAGTCCCCC

        -
1089    GAATCGGAATCTGACCACGAGCACGAGGAGGCAGACCTTGAGCTGTCCCTCCCCGAGTGCCTAGAGTA
           NESP5-------------->
                                                                        <-----
1157    CGAGGAAGAGTTCGACTACGAGACCGAGAGCGAGACCGAGTCCGAAATCGAGTCCGAGACCGACTTCG

        -------NESP7
1225    AGACCGAGCCTGAGACCGCCCCCACCACTGAGCCCGAGACCGAGCCTGAAGACGATCGCGGCCCGGTG
                                                              PvuI

                                                    NESP 6------------>
1293    GTGCCCAAGCACTCCACCTTCGGCCAGTCCCTCACCCAGCGTCTGCACGCTCTCAAGTTGCGAAGCCC

1361    CGACGCCTCCCCAAGTCGCGCGCCGCCCAGCACTCAGGAGCCCCAGAGCCCCAGGGAAGGGGAGGAGC

             <--------------NESP2
1429    TCAAGCCCGAGGACAAAGATCCAAGGGACCCCGAAGAGTCGAAGGAGCCCAAGGAGGAGAAGCAGCGG

1497    CGTCGCTGCAAGCCAAAGAAGCCCACCCGCCGTGACGCGTCCCCGGAGTCCCCTTCCAAAAAGGGACC
                                          MluI
```

```
          <-----------NESP11
1565  CATCCCCATCCGGCGTCACTAATGGAGGACGCCGTCCAGATTCTCCTTGTTTTCATGGATTCAGgttа
                    NESP3--------------->

1633  gttgcccaccgctaaactggggagcctgagggcggtgtgggagcagcgcaggtggaaaggaggtgaga
1701  aggaaaggcaggtcaggggcgagtgggaagagaggaggctcagctggtcagcctgggatcgggggtca
1769  gggtgaggcggcgagggctcccccaaacttcccaggatgccaggggcgccctggtggccaaaggcttg
1837  ttggacggcggggcgcacgccgtgcgtcccgctggagacaacctgaggtctccgagctggtgccccgg

                    NESP10------------>
1905  ctacgcgactgaatcccgaacgcaccccaatttggagctgcacacccaaacggcattggtaagtcact
                    <----------NESP 12

1973  tgttttgcgcgcttttcttcctcctagaaagactagtctcaaataagttggccttctcaggtgtccaa

                                        <-----------NESP26
2041  aatgtggttcggaggtgcgcgcgccaactttcacgatgtgagagcagccgcgctgtagagacaccgtt

2109  gaaatgtgcggaaagtaatctgaatgggaatgggcgagaactctagagactgaccacccgggagggaa
2177  gtcacgcgcgcgcggcgcctaagcagctcagagccggagcccaggtcccagagctgacaattaagccg
                                                                  SacII

2245  cgggacctccgcgccagtgcctccagctgccgtgcgccagccttggccgccacagcccgcctcccgtc
2313  gctcgcgggacagagaccgcctcaaagagcgtgcgcaacctgcccgcgcgcgccggagctgacctctcc
                                    FspI

          NESP14------------>
2381  cggcggcgggcggttaggggaaagtacctgggggaaaggtagaggaggtaaggggacccttggggatg
          <-----------NESP13

2449  cccctacgggctaccagggttgaacgcacaggcatggtcacgtcggggtattgccaagcttttggcgc
                                                          HindIII
2517  agctggtcgggtggccaggctgcatgcggcttagcaggagacgtcctgggctgtttgcgcaggacctc

                                                                    -----
2585  tggaggccctcgagatcgtcgcaagtggaaaggtaaagcggaacaagggacaggctggagacgggggt
            XhoI
                                                                    <---
      -NESP 17----->
2653  cgcgtctaacatcaggataacttacAATTCGTTTCCAAAGAGCGCGGTACGCGCAGAGCTGGGGAAAG
      -NESP 18------          AS-EXON IV

2721  GTGTTGGATCCGGCGCCAGTCCTTGGACGATCAGTCGTCGAAAAGGAGGCAGGGGCCGGCGGCTCCGG
           BamHI                                                NgoMI

2789  CAGCCTTGGGGAGGGGAGGCGAGGCGGCGTGCTCCGGCCATTTTCAGCACGGGTAGAGTTAAGTTTAA
2857  TCTGTAACCTGGAGGGGAATCTGCTCTGATGACCCAGCACAAAAACGGCAGCAATCTGGTAACGCACC

2925  TTCGGAAGGGAAGGCGTGCTCTCCTCGCCAGTCCTGGGGGGAAAGActgataagtgaaatgtctttaa

      NESP21------------>
2993  aaaatctcatccgacttggaaattgatttttttttcctggtgtgcgtgtctcattttgcgccggcgctt
      <------------NESP22                                          NgoMI

                                                  <------------NESP25
3061  aattatacggatgacaacgatttggaggaatagtaaagcgttttgagggcttgagatgaatgacggtc

3129  aggttcttaactcctaagaaaaggatgatgatagaaggaaaataccttaaatggtcttccttccagga
3197  cctattccgtaggaaaggcgtcgcggcgcccaagacttagttctgccgggctgcagcgcagactgtgg
3265  ttggatgcttttgtggtcttccctagtaaatacggagaaactagtttatttgcaagtggttttctgcg
3333  acagcaatgttagtctcagaagtttggaggctgcggaagggctaggcgtgccccgcttttggcttgcc
```

```
3401   cctaagtcgaaattcctgcttaatgtaaacaatgatggtgtgttggctttctgtgacattaagtgttc

                              NESP23------------>
3469   agtttttcctctggtaatgtctcagcctgagttcaccaagcaaaatgtctctagggacaagcaagaaa
                              <-----------NESP24

3537   aactagggattgtgttgaatttctttggtgtgcctttcaagtgtgggagctatcttcagactgggtat
3605   tccgggagacccgtttttcccacagagaattattcattagagaaatgagcatgttgccttgttgcagg
3673   atgaataaacaagtttttttaaaagccctccgtgtcgagcccacacctcacacccatcttgctgttttt

                                 NESP15---------------->
3741   taaacagaagcccaacatagggatgtttcctgaactgtaagcaatgtaccatgfttcacatgtagcga
                              <------------------F8F2


3809   ggaggggcaatctaggaacacacactgtctttcagagtcttccacaaactgcctcgccatctgcatta
3877   cagcccaattgtaaaaacctgttacctaccggcacacactctcccccgctcaattcccctcccccgct
3945   caattctcctcccccgctcaattcccctcccccagcttccctctttacaccctgattcaagggtggaa


                                          <------------NESP16
4013   aagccagtgcagaccatatgaagcaactgagagctataccaacccccgaggaaagcacctatggccct
                                          F8BglM13F1--------->
4081   ccctgagaagacaggattttgctatatagacttcgcaaaccttctctctgcctgagagactggtactc
4149   agctccttcctttccagagccatttttagcaacttccccaagcattcccaactcactagtctcccag
4217   ctggcccttgccccaccctgctcaactggggttagcatttgaataggccgacttgccccttttacata
4285   ggtgaggtctgcggggtgaggataggctcccaagaaaaagc***agatct.***
                                             *Bgl*II
```

3.3.7.3 Predicted NESP55 protein sequence analysis

The predicted translation product of the NESP55 cDNA has a molecular mass of 28,031 Da (Figure 3.18). The sequence of bovine NESP55 had already been determined (Ischia et al., 1997) but for further comparative analysis the murine NESP55 amino acid sequence was also determined. (Dr. B.E. Hayward and Professor Bonthron). This was done by BLAST searching, identifying an IMAGE cDNA clone (746837) that contained part of the murine NESP55 homologue, after which the complete sequence of murine NESP55 was determined from this cDNA clone and from RT-PCR products derived from a βTC3 pancreatic islet-derived cell line.

The predicted human NESP55 translation product has 245 amino acid residues, compared to 241 amino acids for bovine NESP55 and 253 amino acid for murine NESP55. Human NESP55 is rich in glutamic acid (15%), arginine (11%) and proline (11%). The C-terminal two-thirds of the protein (amino acids 75-245) contains a number of repetitive motifs. One of the repetitive regions lies at the middle of the predicted protein (amino acid residues 102-143). This sequence is a dipeptide repeat in which every second residue is glutamic acid (E), of the form E-(T/S/P). Most of the differences between human, bovine and murine NESP55 sequence occur within these repetitive regions, whereas the first 50 amino acids are highly conserved (Figure 3.18).

Analysis of the codon preference of the human NESP55 sequence suggests that selective pressure is imposed at the DNA and not just at the protein level. Examples of this are that of 37 Glu codons, 30 are GAG and seven are GAA (the proportion expected from human codon preference tables is 21 GAG, 16 GAA). Similarly, another example is that of 14 Thr codons, 12 are ACC, two ACT, 0 ACG, 0 ACA (expected; five ACC, three ACT, two ACG, four ACA) (as deduced using the protein sequencing analysis software at the HGMP MRC UK bioinformatics site).

```
          1                                                  50
bovnesp   mdrrsrpqlg rrarhnyndl cppigrraat allwlscsia llralatsst
humnesp   mdrrsraqqw rrarhnyndl cppigrraat allwlscsia llralatsna
musnesp   mdrrsraqqw rrarhnyndl cppigrraat allwlscsia llralassna

          51                                                 100
bovnesp   raqqraaa.q rrtflnahhr s.....aaqv fpeppes... .dhedtdfep
humnesp   raqqraaaqq rrsflnahhr s.....gaqv fpespesesd heheeadlel
musnesp   raqqraa..q rrsflnahhr saaaaaaaqv lpessesesd heheevepel

          101                                                150
bovnesp   slpecpeyqe eefdyesete s..eseiese tefetesdta pttepetepe
humnesp   slpecley.e eefdyetese t..eseiese tdfetepeta pttepetepe
musnesp   arpecleydq ddyetetdse tepesdiese teietepete petepetepe

          151                                                200
bovnesp   depgpvvpkr ptfhqslter lsalrlrspd aspsrappst qesesprqge
humnesp   ddrgpvvpkh stfgqsltqr lhalklrspd aspsrappst qepqsprege
musnesp   derg...prg atfnqsltqr lhalklqsad asprraqptt qepesasege

          201                                                250
bovnesp   e....pedkd prdpeeseep keeekqqqhr ckpkkpt.rr dpspespskr
humnesp   elk..pedkd prdpeeskep kee..kqrrr ckpkkpt.rr daspespskk
musnesp   epqrgpldqd prdpeeepee rkeenrqprr cktrrparrr dqspespprk

          251
bovnesp   gaipirrh
humnesp   gpipirrh
musnesp   gpipirrh
```

Figure 3.18. Multiple sequence alignment of bovine (bov), human (hum), and mouse (mus) NESP55 amino acid sequences. The alignment was obtained by the analysis of the sequences with the Pileup program. Residues in common between two or more sequences are shaded (modified from Hayward et al., 1998b).

3.3.7.4 Allele-specific methylation in the region of the NESP55 exon

To assess the methylation status of the region containing the NESP55 exon, its restriction map was referred to. The subclone Bgl34 was then digested with the enzymes *Kpn*I and *Fsp*I. The resulting 2.6 kb fragment (as marked in Figure 3.19) was used as probe in a Southern blotting analysis (by Dr. L. Strain). This probe was predicted to detect fragments containing a number of rare-cutting methylation-sensitive restriction sites (*Xho*I, *Ngo*MI, *Mlu*I, *Fsp*I and *Sac*II). Parthenogenetic F.D. or normal female control lymphoblastoid cell line DNA was double-digested with *Bgl*II and one of the methylation-sensitive enzymes. The analysis showed that for all of these enzymes, in normal controls the *Bgl*II fragment is partially resistant to cleavage, indicating methylation of these sites on at least one allele (paternal). Meanwhile in the parthenogenetic DNA all of the enzymes but *Mlu*I cleaved the *Bgl*II fragment to completion. Therefore, these restriction sites are unmethylated on the maternal allele, and methylated on the paternal one (Figure 3.19). It may be noted that although most of the differentially methylated region lies within the NESP55 exon, one of the paternally methylated *Ngo*MI sites lies within an exon of the paternally expressed antisense transcript.

Figure 3.19. Methylation analysis of NESP55. Restriction sites whose methylation status was examined are indicated above by – or + to indicate unmethylated or methylated status, respectively, on maternal and paternal alleles. The hatched bar indicates the probe used in Southern blotting in the present study. The methylation indicated for the XLαs exon and exon 1 is as described elsewhere (this chapter section 3.3.4). The restriction enzymes used for the analysis are represented by letters as described in Figure 3.11 and 3.16. The *Ngo*MI site situated within exon IV of the antisense transcript is in bold letters (modified from Hayward et al., 1998b).

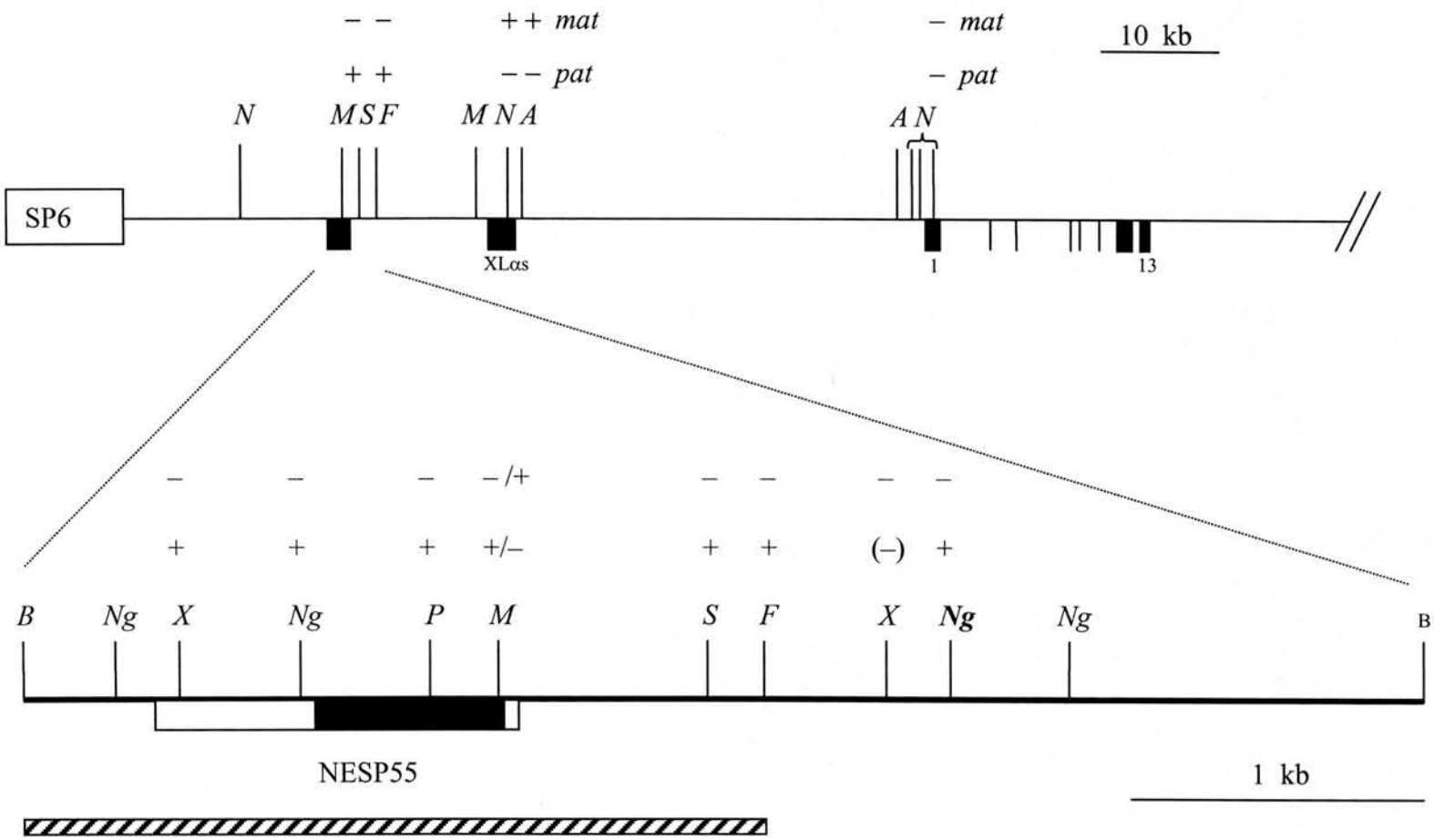


Figure 3.19. Methylation analysis of NESP55

3.3.7.5 NESP55 is expressed only from the maternal allele

As NESP55 and XLαs exons were found to have an opposite methylation pattern, it was also further investigated whether they also show opposite patterns of allele-specific expression. For this purpose, RT-PCR analysis (by Dr. Hayward) was again performed using fetal tissues known to be informative for the *Fok*I polymorphism in exon 5.

The RT-PCR was made specific for NESP55 using a unique upstream primer. From a total of 19 tissue samples from 6 informative fetuses, uniform results were obtained in several tissues. In each case NESP55 transcripts were exclusively derived from the maternal allele.

3.3.8 ANTISENSE TRANSCRIPT AT THE *GNAS1* LOCUS

Once the complete sequence of the Bgl34 subclone had been obtained, the whole sequence was used for further database searches using BLAST. Besides the NESP55 cDNA sequences identified as matching region ***(ii)*** of the genomic clone, region ***(iii)*** of the sequence, as alluded to above, also identified another distinct EST (Acc. number T84962). This EST had in fact also already been identified during the analysis of another genomic subclone derived from PAC dJ309f20 (the already mentioned BamC6, lying further downstream towards XLαs). The IMAGE clone 111764 (named 088f21 by the UK HGMP Resource Centre) corresponding to this EST was then completely sequenced (as described in Chapter 4).

The same transcript was again later identified by further BLAST searching, using the complete sequence, later obtained in our laboratory, of the region between NESP55 and XLαs in human and mouse. Further RT-PCR and 5′-RACE experiments indicated that the cDNA 111764 was part of a bigger transcript composed of 5 exons, and 1.1 kb in total length. Comparison of the cDNA sequence to the genomic sequence showed that this transcript was spliced, and must have been transcribed in the antisense direction relative to NESP55. Furthermore, it was also later shown that

this antisense transcript is imprinted, being expressed from the paternal chromosome only (Hayward and Bonthron 2000).

3.4 DISCUSSION

In the present chapter, I have described studies carried out on the *GNAS1* gene as part of a project aimed at elucidating new aspects of its genomic structure and its complex pattern of imprinting. It became clear that the original description of the *GNAS1* gene structure (Kozasa et al., 1988) was incomplete. Furthermore, the new information obtained during this study has gone some way towards resolving the confusion over the imprinting status of *GNAS1*. In particular, it is now clear that the conclusions drawn from earlier studies of *GNAS1* expression were somewhat misleading, since these studies (Campbell et al., 1994, Williamson et al., 1996) did not attempt to distinguish the expression of the differently imprinted *GNAS1* transcripts.

During the course of the present studies, new exons of *GNAS1* were defined, and it was also shown that the promoters upstream of these novel exons display a complex pattern of allele-specific expression. The new, extended *GNAS1* gene encodes three different products:

(i) the α subunit of the stimulatory G protein G_s; this is the major product in most tissues, and was shown by RT-PCR specific for exon 1 to be biallelically derived in several human fetal tissues. This fact explains the biallelic expression previously observed using an RT-PCR assay that did not distinguish the different *GNAS1* gene products (Campbell et al., 1994).

(ii) NESP55, a neuroendocrine secretory granule protein, expressed exclusively from the maternal allele.

(iii) XLαs, a plasma membrane associated, adenylate cyclase-activating G_s variant protein, that is expressed exclusively from the paternal allele.

Both NESP55 and XLαs display a similar pattern of tissue distribution, being expressed at highest level in neuroendocrine tissues including the pituitary, the adrenal medulla, and certain parts of the brain. However, RT-PCR analysis showed that these proteins are strictly monoallelically derived, even in tissues with only a very low level of expression (Hayward et al., 1998a, b).

In addition to these protein-coding transcripts, a novel non-coding antisense transcript was later identified at the *GNAS1* locus. This spliced transcript originates ~2 kb upstream of XLαs, and terminates ~20 kb upstream of NESP55. It is expressed from the paternal allele, and has been hypothesized to have a role in suppressing the paternal NESP55 promoter by some cis-acting mechanism (Hayward and Bonthron, 2000). A similar transcript has also been identified in the mouse *Gnas* gene (Wroe et al., 2000) and sequence comparisons between human and mouse show the antisense promoter and first exon to be highly conserved, further suggestive of an important functional role for the antisense trancript (Hayward and Bonthron, 2000).

Although, as mentioned above, $G_s\alpha$-encoding transcripts are predominantly biallelic, it does appear that there is predominantly monoallelic expression in certain cell types. In our laboratory, for example, it has been recently demonstrated that $G_s\alpha$ is predominantly maternally derived in the human adult pituitary. This correlates with a strong maternal bias in the allelic origin of the activating *gsp* mutations of codon 201 and codon 227 in growth hormone-secreting pituitary adenomas; 21 out of 22 such mutations were found to occur on the maternal allele (Hayward et al., 2001).

3.4.1 *GNAS1* GENE STRUCTURE

As originally described by Kozasa et al. (1988), the human *GNAS1* gene comprised 13 exons and 12 introns, spanning ~20 kb in 20q13. However, further studies carried out by a number of different research groups later identified new exons of this gene. The first new exon described was named exon A (also referred to as 1A in the mouse; Liu et al., 2000a) and is located 2.5 kb 5′ to the original exon 1 (Ishikawa et al., 1990, Swaroop et al., 1991).

The remaining novel *GNAS1* exons were identified as described in this thesis. All but one of these novel exons (the exception being exon 3N, that is located in the intron between the originally described exons 3 and 4) are located at the 5′ end of *GNAS1*, in the order: NESP55-XLαs-A20-A21. NESP55, the most 5′ of these exons, is located ~ 49 kb upstream of the original exon 1. The XLαs, A20 and A21 exons are located ~ 35 kb from exon 1 (Figure 3.20).

Like exon 1 and exon 1A, these new exons at the 5′ end of *GNAS1* are alternatively spliced onto exon 2 of this gene. Although the original XLαs-containing IMAGE cDNA clone contained also exons A20 and A21, RT-PCR experiments indicate that most XLαs-containing transcripts do not contain the A20/A21 exons. Furthermore, as A20/A21-containing transcripts do not have a continuous open reading frame, they may be unimportant splice forms. Like the XLαs exon, the NESP55 exon, too, is spliced onto *GNAS1* exons 2-13. However, unlike XLαs, the open reading frame of NESP55 is contained entirely within its own exon, so that the downstream splicing events should not affect the function of the encoded protein. NESP55 transcripts terminating in exon 3N should therefore be functional (whereas XLαs transcripts that use this alternative terminal exon presumably are not).

Figure 3.20. Genomic structure of the *GNAS1* gene.

A) Map (to scale) obtained from the analysis of PAC clone dJ309f20. (Left) The SP6 end of the PAC insert and NESP55 and XLαs exons. (Right) Exons 1-13 of *GNAS1* (as originally described by Kozasa et al., 1988) are indicated by an inverted brace. The exon A has been described by other groups (Swaroop et al., 1991, Liu et al., 2000a).

B) Map (not to scale) showing the organisation of the XLαs region of *GNAS1* and the new exons identified. Some of the restriction sites referred to in the text are shown for reference: *S*, *Sac*I; *B*, *Bam*HI; *As*, *Asc*I. The distance between the two *Asc*I sites is ~ 30kb. The 2479 bp genomic region that was sequenced is indicated between the *S* and *B* sites. The exon A20 is in black; other exons are grey or white for untranslated regions (modified from Hayward et al., 1998a, b).

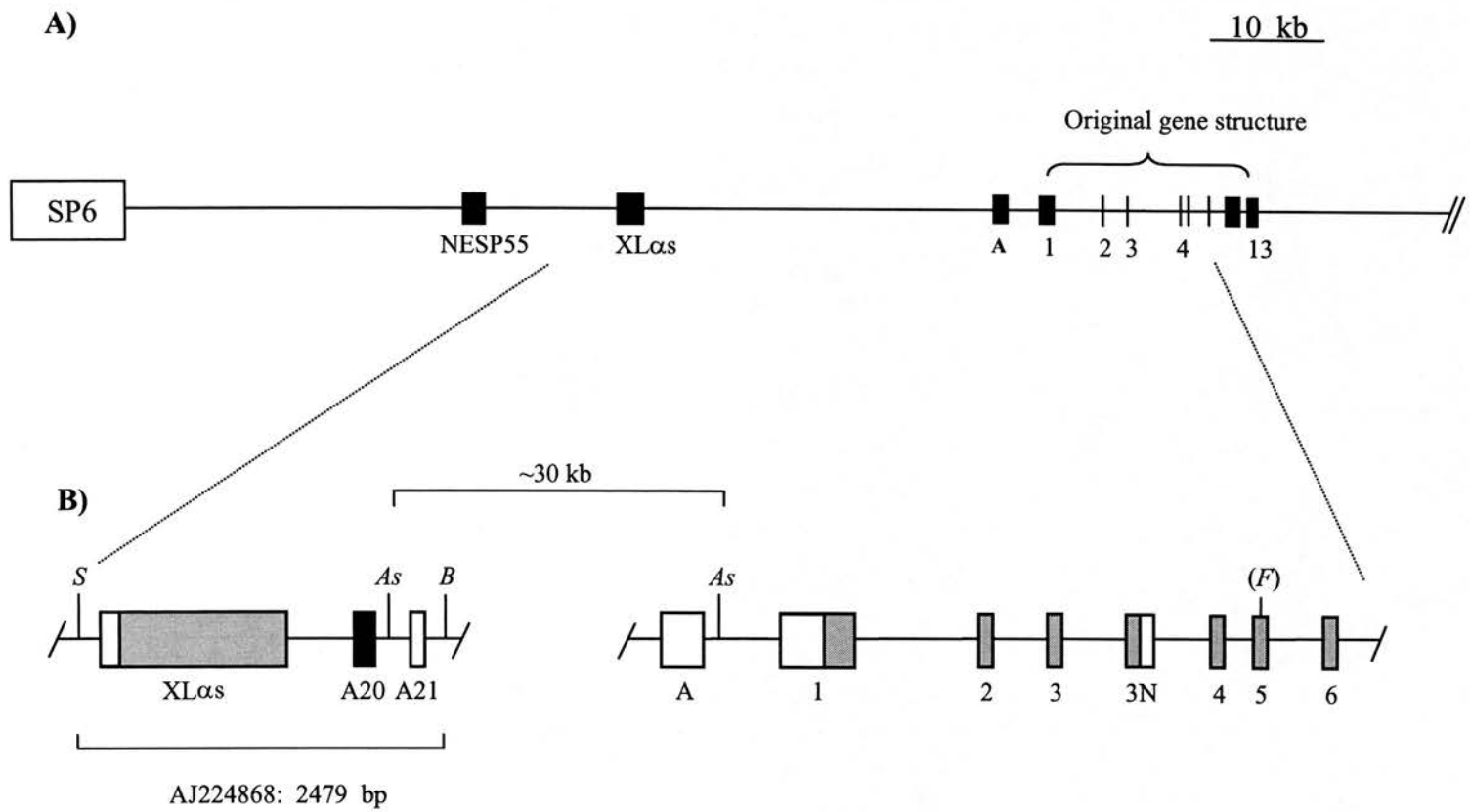


Figure 3.20. Genomic structure of the *GNAS1* gene

3.4.2 XLαs EXON

When the translated open reading frame of human XLαs was compared with the corresponding predicted rat protein sequence, it was noted that there are well-conserved regions at the N- and C- terminal extremes of the XLαs segments. The central, repetitive part of the protein sequence, however, is much less well conserved, although qualitatively similar features do appear to be present in both species. As mentioned previously, the rat XLαs protein has a tetrapeptide repeat, based on EPAA, while the predicted human protein has a tripeptide motif that is again rich in proline and alanine and in an acidic amino acid residue (aspartate rather than glutamate as in the rat protein).

The precise significance of such repeats is not clear, although they have been noted to occur frequently in imprinted genes (Neumann et al., 1995a, b). There are some interesting parallels between the XLαs repeats and those within the $p57^{KIP2}$ protein encoded by the imprinted gene *CDKN1C* (11p15.5). In the mouse, $p57^{KIP2}$ has four structurally distinct domains: an amino terminal CDK (cyclin-dependent kinase) inhibitory domain, a proline-rich domain (25% proline), an acidic repeat domain and a carboxy-terminal domain. In humans, although a repetitive proline-rich domain is again present, it is not well conserved in sequence, consisting in this case of a PAPA (proline-alanine) repeat (Matsuoka et al., 1995).

In the rat XLαs protein, the repeat domains form two of a total of four domains that have been defined. These are: an EPAA repeat domain (residues 37-104), an ARAA repeat domain (103-187), a proline-rich domain (205-228) and a cysteine-rich domain (237-304) (Ugur and Jones, 2000). XLαs has been proposed to comprise two functionally independent units that are connected by the proline rich region. The cysteine-rich region (which is palmitoylated and contributes to the membrane attachment of XLαs), the C-terminal region of the XL domain (which is implicated in the βγ-subunits binding) and the αs domain are all suggested to belong to one unit that exerts all $G_s\alpha$-type functions of XLαs. The other unit, whose function is as yet

unknown, comprises the alanine-rich repetitive region of the XL domain (Klemke et al., 2000, Pasolli et al., 2000).

As mentioned, XLαs behaves in many but not all ways in a functionally similar fashion to $G_s\alpha$: it forms a heterotrimer with βγ-subunits, binds GTP and undergoes a conformational change upon GTP binding. In the GTP-bound state, it also activates adenylate cyclase. However, XLαs does not appear to be activated by known $G_s\alpha$-coupled receptors (Pasolli et al., 2000).

Immunofluorescence and subcellular subfraction analyses demonstrated that XLαs and its truncated splice variant referred to as XLαN1 (derived from use of exon 3N) are neuroendocrine-specific proteins expressed in many but not all, neuroendocrine cells (Klemke et al., 2000). The main cellular localisation of XLαs is to the plasma membrane (Klemke et al., 2000), with a minor proportion being associated with the Golgi complex (Ugur and Jones, 2000, Klemke et al., 2000).

As described above, the demonstration that transcription of the XLαs mRNA was exclusively from the paternal allele represented the first molecular confirmation that *GNAS1* was indeed an imprinted gene. Nonetheless, the identification of a transcript that was paternally expressed was puzzling; it had been expected, given the pattern of inheritance recognized in PHP-Ia families, that a maternally, rather than a paternally expressed transcript, might account for this parent-of-origin dependence.

If XLαs were to be implicated in the pathogenesis of PHP/PPHP, its paternal monoallelic expression would demand consideration of the following issues: Firstly, how could a paternally inherited mutation (presumably resulting in reduced function of both $G_s\alpha$ and XLαs) have a less severe effect than a maternal one, that presumably affected $G_s\alpha$ but not XLαs?. Secondly, it had been previously shown in the mouse that the *Gnas1* paternal allele was preferentially expressed in the renal glomerulus (Williamson et al., 1996). It was not clear how or if this reflected imprinting of the XLαs transcript. In contrast, at least at the whole organ level, *GNAS1* in humans showed no evidence for tissue-specific imprinting (at least in the

fetal tissues examined) (Campbell et al., 1994). However, these earlier studies had not distinguished transcripts derived from the various different *GNAS1* promoters. Because of the inability to reconcile all these different pieces of evidence, it still seemed at least possible that there might be yet other classes of transcript derived from *GNAS1* that were exclusively maternally rather than paternally derived, at least in some tissues. A systematic search for *GNAS1* transcripts with alternative 5′ ends resulted in the identification of another novel 5′ exon, encoding NESP55.

3.4.3 NESP55 EXON

NESP55, the neuroendocrine secretory protein 55000, was first identified fortuitously, during screening of cDNA libraries from bovine chromaffin cells with an antibody against secretogranin II. NESP55 belongs to the group of acidic proteins known as chromogranins. Chromogranins are expressed in neuronal and endocrine tissues, and secreted from large dense core vesicles. They have between 200 and 700 amino acid residues, with glutamic acid as their most abundant amino-acid residue. The chromogranins are classified into five groups: chromogranins A and B, secretogranins I and II, and the so-called VGF group. Within their primary amino acid sequences, several pairs of basic amino acids are found that are used as cleavage sites for endopeptidases. An example is the C-terminal octapeptide GAIPIRRH isolated from bovine chromaffin granules and from adrenal gland where 40% of NESP55 is processed to the peptide GAIPIRRH (Bauer et al., 1999).

The evidence that supports the classification of NESP55 as a member of the chromogranins is as follows: like other chromogranins NESP55 is characterised by a high abundance of acidic amino acids (21%); it is heat stable, and it is localised to the soluble content of the large dense core vesicles of the adrenal medulla, brain and endocrine tissues (Ischia et al., 1997). As mentioned above, cDNA clones containing the NESP55 sequence were first isolated from bovine chromaffin cell libraries. One of these clones was 1499 nt in length with an open reading frame coding for a protein of 241 amino acids with a predicted molecular mass of 27, 494 D (Ischia et al., 1997).

NESP55 was localised within the cell to the large dense secretory vesicles, and in terms of its tissue distribution, is most highly expressed in the adrenal medulla, in the anterior and posterior pituitary and in various regions of the brain. In the rat, NESP55 was not detected in any of the cortical areas of the brain, but it was present in the basal ganglia, hypothalamus and midbrain, particularly those areas rich in noradrenergic, adrenergic and serotonergic neurons (Bauer et al., 1999).

In our studies on human NESP55, we assigned the initiation codon as the first ATG occurring after an in-frame upstream stop codon at nt 685 of our sequence. This indicated that the predicted NESP55 protein product is 245 amino acid residues in length (vs 241 for bovine and 253 residues for murine NESP55), with a predicted molecular mass of 28, 031 Da. Human NESP55 is rich in glutamic acid, arginine and proline. The C-terminal two thirds of the protein (including amino acids 75-245) contains a number of repetitive motifs, and it is within this region that most of the differences between human, bovine and murine NESP55 occur.

It is noteworthy that on sucrose gradient centrifugation, NESP55 was found to copurify with secretogranin II, as had also been reported for XLαs, and also that both NESP55 and XLαs have a similar tissue distribution. Both of these proteins are encoded by the *GNAS1* gene, and yet they do not share any protein-coding sequences, since their open reading frames do not overlap. Furthermore, their respective promoters are oppositely imprinted.

One additional consideration that results from the opposite imprinting patterns of the overlapping NESP55 and XLαs transcripts relates to splicing specificity. In particular, it would appear necessary to propose that the splicing of NESP55 to exon 2 must be strongly dominant over XLαs splicing to the same acceptor site. This conclusion can be reached because the maternally derived nascent transcripts from the NESP55 promoter must contain all the exons NESP55, XLαs and 1-13. If any splicing of XLαs to exon 2 occurred in these transcripts, it would show up in RT-PCR experiments as maternally expressed XLαs mRNA, and yet none is identifiable. Possibly therefore, the splicing of the NESP55 transcript occurs in a strictly

directional 5′ to 3′ fashion, whereby selection of the most upstream splice donor site by some form of "scanning" mechanism results in removal of the XLαs exon during the first splicing event, before any XLαs-exon 2 splicing can occur.

One additional unusual feature of the NESP55 mRNA is that exons 2-13 do not form any part of its coding region, but rather are entirely contained within the 3′ untranslated region of the mRNA. The presence of multiple introns within 3′ untranslated regions is otherwise unusual in mammalian genes.

Bovine NESP55 appears to be proteolytically processed within the chromaffin granule to smaller peptides that have been proposed to be physiologically active. One peptide in particular, comprising the residues Leu-Ser-Ala-Leu (LSAL), a tetrapeptide sequence which has been named 5-hydroxytryptamine-moduline and is flanked by arginine residues, was thought suitable for cleavage by prohormone convertases. LSAL is an endogenous antagonist of the serotonergic 5-hydroxytryptamine IB receptor subtype, and (before the sequence of human NESP55 described here was known) it had been considered that the LSAL peptide in humans might contribute to the pathophysiology of serotonergic transmission disorders (Ischia et al., 1997, Kim et al., 2000).

However, in our human NESP55 sequence, we have found that the amino acid sequence corresponding to the LSAL tetrapeptide in bovine NESP55 is not conserved, being LHAL in humans (as is also the case in mouse). It is therefore far from clear whether NESP55 is the physiological precursor for the LSAL peptide, which also calls into question its postulated involvement in modulating serotonergic transmission. One genetic association study has been performed to address the proposed involvement of NESP55 in patients suffering from autistic and obsessive compulsive disorders (Kim et al., 2000). In this study, six polymorphisms that do not change the amino acid sequence were identified, as well as a deletion of 24 bp in the coding region that results in the in-frame deletion of 8 amino acid residues. In the family under analysis (suffering from obsessive compulsive disorder) only heterozygotes for the deletion were identified, they got the mutated alleles from their

mothers and had a variable clinical presentation with no specific symptoms or phenotype and even one clinically normal individual was described.

The finding of a phenotypically normal subject who has a mutated allele of maternal origin may suggests the possibility that the deletion does not critically compromise NESP55 function or that it may have a supplementary role as a risk factor (Kim et al., 2000).

3.4.4 ANTISENSE TRANSCRIPTS AND REGULATION OF IMPRINTING OF THE *GNAS1* LOCUS

Only 14 kb separates the maternally active NESP55 promoter from the exclusively paternally active promoter that directs the synthesis of XLαs. The proximity of these promoters suggests the possibility of co-ordinate regulation in a reciprocally, mutually exclusive allele-specific manner. One hypothesis to explain how NESP55 and XLαs promoters may regulate each other invokes mutual suppression in *cis*. This might be exerted as a consequence of transcription originating at one promoter and extending across the other promoter region. Such a mechanism has been shown to be involved in regulating the maternally expressed *Igf2r* gene. In this system, a paternally expressed antisense transcript originates from a differentially methylated region within intron 2 of the gene and is transcribed across the upstream sense *Igf2r* promoter. Deletion of the intronic DMR results in loss of the antisense transcript and biallelic transcription from the sense promoter, providing direct evidence for a *cis*-effect of the antisense transcript (Wutz et al., 1997).

A specific search for transcripts within the NESP55-XLαs genomic interval (Hayward and Bonthron, 2000) resulted in the demonstration that the maternally methylated region 2-3 kb upstream of the XLαs exon gives rise to a spliced polyadenylated antisense transcript which spans the upstream NESP55 region. This antisense transcript was shown to be imprinted and expressed only from the paternal allele (Hayward and Bonthron, 2000) (Figure 3.21).

Figure 3.21. Map of the region of the *GNAS1* locus containing the antisense exons. A) The exons NESP55, XLαs and the originally described $G_s\alpha$ exons 1 to 13 are shown for orientation (map to scale). B) The antisense exons (roman numerals) are shown as open rectangles and the NESP55 and XLαs exons as filled rectangles. The arrows show the direction of the transcription and the respective active chromosome is indicated by *pat* (paternal-specific transcription) or *mat* (maternal-specific transcription). Some of the splicing patterns observed for the antisense transcript are shown. The *Not*I restriction site is indicated (modified from Hayward and Bonthron, 2000).

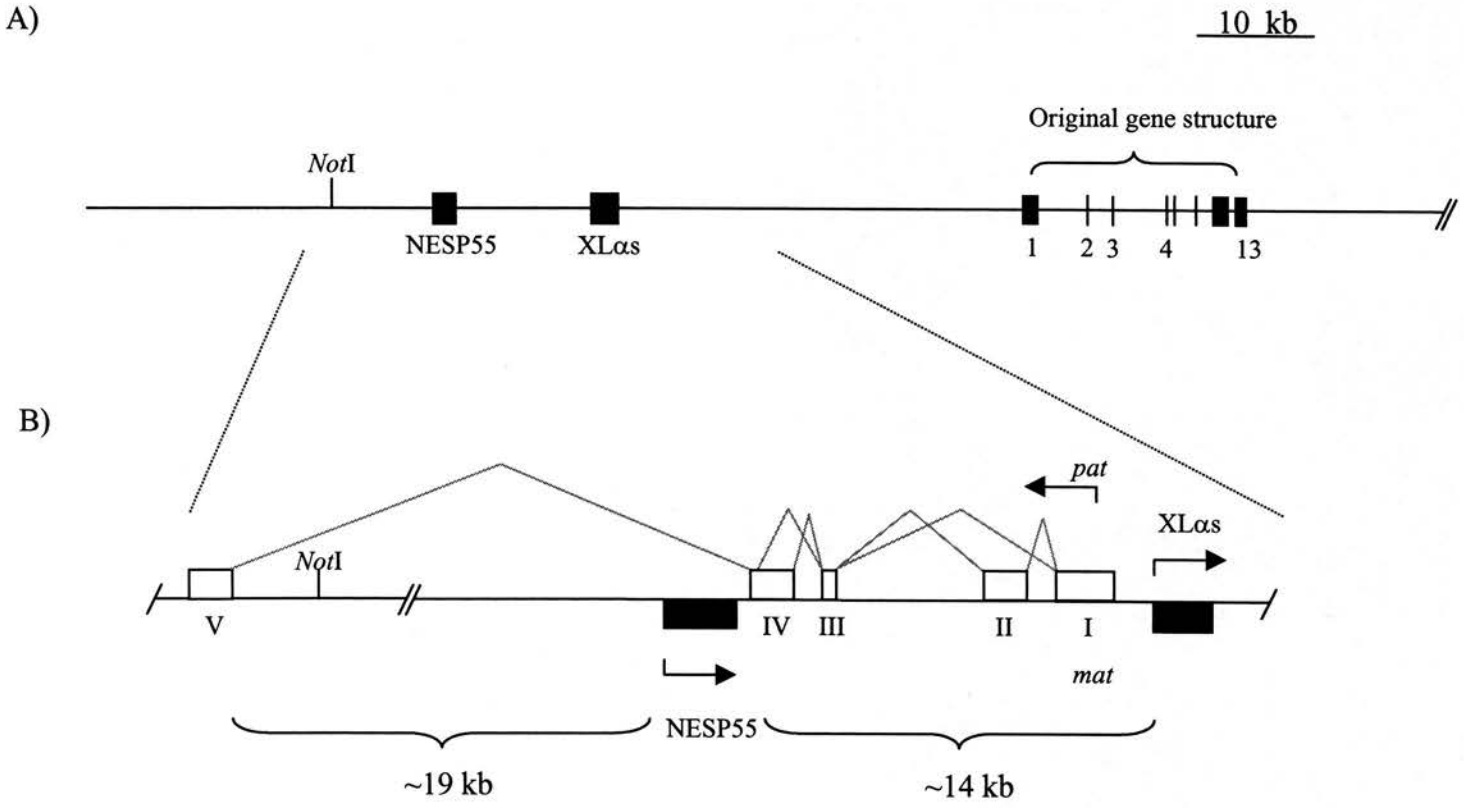


Figure 3.21. *GNAS1* antisense transcript

The mechanism of action of such putatively regulatory antisense transcripts has not been defined. However, the importance of RNA as a regulatory molecule for imprinting may lie in its ability to act genetically in *cis*, unlike proteins which must be translated in the cytoplasm and then reimported into the nucleus before acting. This kind of direct link between imprinting and *cis*-inactivation has also been studied in the context of X-inactivation. In mammals, X inactivation requires the action of a control switch called the "X inactivation centre" (XIC). The *cis*-acting silencing function of the XIC requires the action of *Xist*, an XIC gene that produces an untranslated RNA which appears to "coat" the inactive X (Lee and Lu, 1999).

In addition to *Xist*, an antisense gene named *Tsix* (downstream of *Xist*) has also been identified within the XIC, and is thought to act as an *Xist* regulator. Targetted deletions of *Tsix* in female and male mouse cells have shown that despite a deficiency of *Tsix* RNA, female cells still inactivate one X, while in male cells X inactivation is blocked. Thus, X chromosome counting remains intact in the absence of *Tsix*. However, heterozygous female cells showed skewed *Xist* expression and primary non-random inactivation of the mutant X. Therefore, it appears that *Tsix* regulates *Xist* in *cis* and determines X chromosome choice without affecting silencing (Lee and Lu, 1999).

Antisense transcripts have been described to be associated with a number of other imprinted genes, including for example: human *KVLQT1* and *UBE3A*, and mouse *Igf2r* and *Igf2*. *KVLQT1* is a potassium channel gene, mutations within which cause the long QT syndrome. The *KVLQT1* gene also encompasses a cluster of Beckwith-Wiedemann syndrome (BWS) chromosomal rearrangement breakpoints in 11p15.5. As discussed in the identification of a paternally expressed transcript in an antisense orientation within *KVLQT1* has revealed that abnormal imprinting of this transcript, named *LIT1*, occurs in some cases of BWS (Figure 3.22) (Lee MP et al., 1999a, Mitsuya et al., 1999).

Figure 3.22. *KVLQT1* antisense transcript. The diagram shows the exon-intron structure of *KVLQT1* and the position of the intronic, differentially methylated CpG island *KVDMR1*. The directions of transcription of the maternally expressed *KVLQT1* gene and of the paternally expressed antisense transcript *KVLQT1-AS* (or *LIT1*) are indicated by a grey and a dark arrow, respectively. Map to scale (modified from Smilinich et al., 1999).

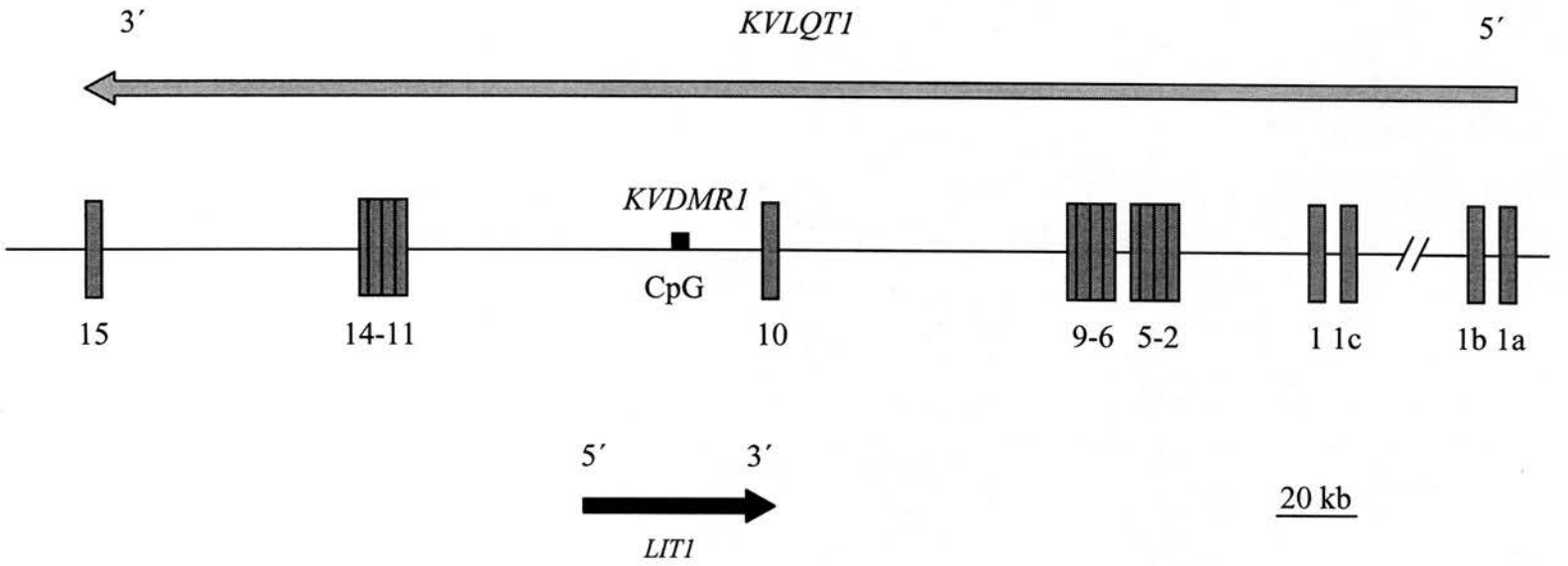


Figure 3.22. *KVLQT1* antisense transcript (*KVLQT1-AS* or *LIT1)*

Methylation analysis revealed that a CpG island associated with *LIT1* was specifically methylated on the silent maternal allele. Furthermore, it was demonstrated that a high proportion of patients with sporadic BWS showed complete loss of maternal methylation at this CpG island, together with biallelic *LIT1* expression. This suggested that the antisense transcript might be involved in the development of BWS. Precisely how this aberration of imprinted antisense transcription results in BWS has not been demonstrated, although it is possible that a pathway involving suppression of the activity of the nearby *CDKN1C* gene is involved (the latter is mutated in a high proportion of familial BWS cases). The mechanism whereby *LIT1* transcription is activated is also not clear. It has been proposed that the *LIT1* CpG island may serve as an insulator between *KVLQTI* and its enhancer; according to this model, methylation of the CpG island would interfere with both insulator function and *LIT1* expression (as has also been suggested for a CpG island near *H19* and *IGF2*). When the CpG island is unmethylated, as it is on the paternal chromosome, it effectively blocks *KVLQT1* from its enhancer, both competing for the enhancer on the same chromosome (Figure 3.22) (Lee MP et al., 1999a, Mitsuya et al., 1999).

In the case of the antisense transcript described at the *UBE3A* locus 15q11-q13, the transcript displays a pattern of imprinting that is opposite to that of *UBE3A*, with preferential expression from the paternal allele in brain. Their location, orientation and expression profile may indicate that the sense and antisense transcripts compete for common regulatory elements. The tissue specificity of the imprinting of *UBE3A* could thus be regulated indirectly by the expression of the antisense transcript. This model would suggest that in tissues where the antisense transcript is monoallelically (paternally) expressed (such as brain), *UBE3A* can only be expressed from the antisense-silent maternal chromosome (Figure 3.23) (Rougeulle et al., 1998).

Figure 3.23. *UBE3A* antisense transcript. Map (to scale) of the 3′ end and the downstream region of the *UBE3A* gene. Coding exons 11, 12, 15 and 16 of *UBE3A* are represented by the dark grey boxes, and the downstream transcript by the narrower marbled grey box. The antisense transcript is represented by the dark box below the *UBE3A* locus. This transcript is intronless, and covers at least the 3′ half of *UBE3A* (from exon 11 to the 3′ end) and additional sequence downstream; its 5′ end is 6.5 kb from the stop codon of *UBE3A*. The arrows indicate the direction of transcription (modified from Rougeulle et al., 1998).

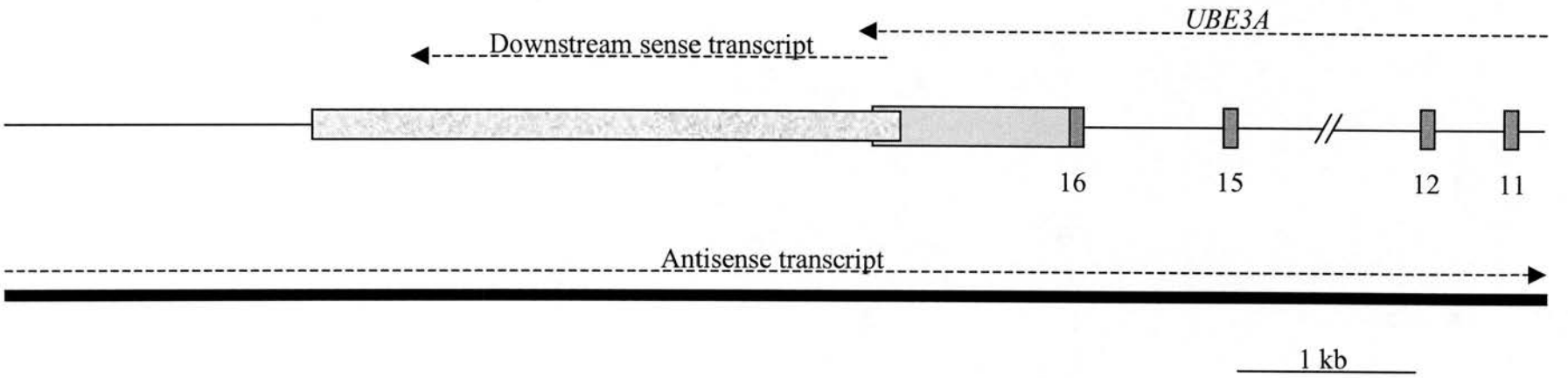


Figure 3.23. *UBE3A* antisense transcript

The non-coding imprinted transcripts seem to be located at least partly within protein-coding imprinted genes (the exception being H19), and are transcribed in the antisense direction relative to the coding transcript. Their sizes range from 1-2 kb to more than 100 kb (Lyle et al., 2000), but a common feature seems to be that they traverse one or more sense exons of the gene with which they are associated. Also noteworthy is the fact that so far all the antisense transcripts described within imprinted genes are expressed from the paternal chromosome, although the small number described so far means that this may be due to chance. As far as the antisense transcript in the upstream region of *GNAS1* is concerned, its importance for the regulation of the locus seems to be indicated by the fact that its putative promoter region is highly conserved in the mouse, presumably due to selective constraints (Hayward and Bonthron, 2000). Aspects of the mouse *Gnas* locus are discussed below.

3.4.5 *GNAS1* LOCUS IN MOUSE

The *Gnas* locus on distal chromosome 2 in the mouse has, like its human counterpart, recently been subjected to detailed scrutiny from an imprinting point of view. Besides the *Nesp* and *Gnasxl* exons, in locations corresponding to those of the human gene, an antisense transcript, named *Nespas*, has been identified. In addition, exon 2 of *Gnas* has been used as the target for an insertional knockout, and observations of this mouse have lead to several new hypotheses regarding the regulation of this locus.

By using representational difference analysis (RDA) to search for regions of differential methylation, Kelsey et al. (1999) identified a sequence that was highly homologous (98.4%) to the bovine cDNA for NESP55. They also identified the murine *Gnasxl* exon (corresponding to human XLαs) in this screen and showed that *Nesp* and *Gnasxl* were exclusively maternally and paternally expressed, respectively, as in humans. As in human, despite their opposite imprinting *Gnasxl* and *Nesp* are part of the same transcription unit as *Gnas,* the *Gnasxl* and *Nesp* transcripts being alternatively spliced onto exon 2 of *Gnas*.

In addition to this imprinting of the upstream promoters, evidence from the *Gnas* knockout suggests that $G_s\alpha$-encoding transcripts of *Gnas* are monoallelically expressed in a tissue-specific manner, in proximal renal tubules (Yu S et al., 1998), and also in brown and white adipose tissue (Yu S et al., 2000). This appears to be the case since in these tissues, the $G_s\alpha$ protein and mRNA levels are much lower in mice in which the null allele is maternal, than in those in which it is paternal.

The *Gnas/GNAS1* locus was the first example of a cluster of imprinted genes in which two oppositely imprinted transcripts share the same exons. As in humans, the possibility was suggested of an expression competition model, in which methylation regulates the availability in *cis* of shared regulatory elements. This could account for the opposite imprinting of *Nesp* and *Gnasxl*, and might furthermore involve the activity of an antisense transcript named *Nespas*, which as in the human gene, was identified in the region between *Gnasxl* and *Nesp* (Li et al., 2000, Wroe et al., 2000).

Specifically, it was postulated that *Nespas* (paternally expressed), may repress the expression of *Nesp* in *cis* from the paternal allele. *Nespas* in that sense could be regarded as an "imprintor" and *Nesp* as the imprinted target. This situation could lead to non-expression of the sense transcript by any of the following mechanisms: occlusion of the sense promoter, inactivation of the paternal allele by localised "heterochromatinisation", or competition for shared transcription factors or enhancers between the sense and antisense promoter. An expression competition model in which methylation regulates the availability of shared regulatory elements also could account for the expression of *Nesp* and lack of expression of *Gnasxl* from the maternal allele (Wroe et al., 2000).

The phenotypes described for mice heterozygous for a *Gnas* knockout denote further effects of imprinting at this locus. In particular, it was observed that when the knockout is inherited paternally, the resulting mice have a hypoactive phenotype resembling that found with the maternal disomy, whereas in contrast when the null allele is maternally inherited, the mice have a hyperactive phenotype similar to that found with the paternal disomy (Yu S et al., 2000).

It has been also observed that the heterozygous knockout mutation in *Gnas* (an insertional disruption of exon 2) leads to opposite effects upon energy metabolism, resulting in obesity when present on the maternal allele (m-/+) and to leanness when present on the paternal allele. The homozygotes (-/-) all die during early postnatal implantation development. It is interesting that like the m-/+ mice, humans with PHP-Ia develop obesity, although the physiological mechanism causing obesity in these patients is not defined yet. Overall, the phenotypes resulting from *Gnas* knockout were felt to support the hypothesis of conflicting parental goals driving the evolution of imprinting, since disruption of the paternal allele leads to decreased prenatal growth, failure to suckle at birth and markedly reduced accumulation of energy store postnatally, whereas disruption of the maternal allele leads to opposite effects (Yu et al., 2000).

More recently, those transcripts originating in the exon 1A region (2-3 kb upstream of exon 1) have been studied in both the mouse and human by Liu et al. (2000b). These authors have shown that exon 1A generates mRNAs only from the paternal allele. Furthermore, this region comprises a new DMR, being methylated exclusively on the maternal allele. The methylation in this region is established during gametogenesis and maintained through pre-and postimplantation development and therefore this upstream region can be considered as a methylation imprint mark. The suggestion was made that the exon 1A region may act as a centre that controls or initiates imprinting of the *Gnas* locus, including the tissue-specific imprinting of the exon 1 transcripts. However, these authors also showed that the exon 1A DMR is methylated in all somatic tissues, while $G_s\alpha$ is only imprinted in some. One possibility is that exon 1A contains binding sites for tissue-specific repressors that can bind to the paternal allele but are unable to bind to the maternal allele because its binding site is methylated. An alternative model would be that exon 1A contains a boundary element that blocks activation of the $G_s\alpha$ promoter by an upstream tissue-specific enhancer on the paternal allele but does not block promoter activation on the maternal allele because it is methylated (Liu et al., 2000a).

Additional evidence suggesting the regulatory importance of the exon 1A region came from analysis of patients with PHP-Ib, in which there is a tissue-restricted hormone resistance, involving the PTH target tissues but not the other endocrine targets that are affected in PHP-Ia. All the PHP-Ib patients examined had abnormal methylation of the exon 1A region, such that both alleles appeared to have a paternal methylation pattern (Liu et al., 2000b). In some but not all patients the abnormal methylation also involved the far upstream NESP/XLαs region. The precise nature of the mutation underlying this epigenetic alteration has not yet been defined. However, it is proposed that the presence of a paternal epigenotype on the maternal allele results in loss of the normally monoallelic expression of $G_s\alpha$ transcripts in the proximal renal tubule. This tissue would be selectively sensitive to this imprinting defect, since in other tissues $G_s\alpha$ transcripts are normally biallelic and hence transcription would not be lost.

Evidence that $G_s\alpha$ transcripts are imprinted in another tissue, the human pituitary gland, has also recently been obtained (Hayward et al., 2001). Using normal pituitaries from individuals heterozygous for the polymorphism in exon 5 of *GNAS1*, RT-PCR products from exon 10 to each of the unique first exons were generated. These RT-PCR products were thus specific for the $G_s\alpha$, XLαs and NESP55 transcripts. It was clear from this analysis that the $G_s\alpha$ transcript in the pituitary, unlike those other tissues that had been previously assessed, has a mono-allelic (maternal) expression.

Relaxation of this imprinting of $G_s\alpha$ was further observed in analyses of growth hormone-secreting pituitary adenomas. While some 40% of such tumours possess *gsp* mutations that constitutively activate $G_s\alpha$ (see section 3.1.7.1.2) 60% of somatotroph adenomas do not possess *gsp* mutations. In normal human pituitary and non-functional adenomas, it was found that $G_s\alpha$ was either expressed exclusively maternally or very strongly biased towards the maternal allele; in contrast in pituitary somatotroph adenomas it was shown that the $G_s\alpha$ expression was often approximately biallelic, indicating that relaxation of the $G_s\alpha$ imprinting had occurred during tumorigenesis (Hayward et al., unpublished results).

As described in this chapter, the regulation of *GNAS1* has proved far more complicated that originally thought, as this gene has been proved not only to be imprinted but also to have a particularly complex pattern of expression, with paternally (XLαs), maternally (NESP55 and $G_s\alpha$- pituitary) and biallelically derived products ($G_s\alpha$ in most tissues). Furthermore, the presence of the possibly regulatory imprinted antisense and exon 1A transcripts adds further complex possibilities to the mechanisms of regulation of this locus. Further experimental studies are needed to assess the specific roles of the antisense transcript and exon DMR A1 in the regulation of the *Gnas1/GNAS1* locus.

The splicing patterns observed for the *GNAS1* transcriptional unit are shown in Figure 3.24.

Figure 3.24. Splicing patterns observed for the *GNAS1* transcriptional unit (not to scale). The *GNAS1* locus on chromosome 20q13 is shown at the centre (**B**) of the figure. The translated regions are shown in grey. The arrows indicate transcriptional active promoters. The figure shows the allele-specific methylation and expression patterns of the transcripts on the maternal (**A**) and paternal (**C**) alleles. The grey lines above and below the exons represent the splicing patterns observed for the region.

A) Maternal allele.- The XLαs and exon A regions are methylated. The maternally expressed NESP55 entire coding region is located in its first exon, the stop codon of the ORF is 5′ to the splice junction with exon 2, so that none of the $G_s\alpha$ coding region is included in NESP55. The asterisk above the arrow corresponding to the transcriptional active promoter of exon 1 is to indicate that although $G_s\alpha$ is biallelically expressed in most tissues, its expression is preferentially maternal in human pituitary gland.

B) ***GNAS1 locus.-*** At the right of the figure the original gene structure of *GNAS1* formed by exons 1 to 13 is shown; the exon A (~2kb 5′ to exon 1) is also shown .The newly described exons NESP55 and XLαs are indicated as dark rectangles. The antisense exons (roman numerals) are shown as open rectangles. The arrow shows the direction of the transcription of the antisense transcript indicated by *pat* (paternal-specific transcription). Some of the splicing patterns observed for the antisense transcript are shown.

C) Paternal allele. On the paternal allele the NESP55 region is methylated, and there is transcription starting at the promoters of exons XLαs, A and 1. XLαs and $G_s\alpha$ proteins have identical COOH-terminal domains, while their unique NH_2-terminal domains are encoded within their respective first exons. The information in the figure is derived and/or modified from Kozasa et al., 1988, Ishikawa et al., 1990, Swaroop et al., 1991, Hayward et al., 1998a and b, Hayward and Bonthron, 2000, Liu et al., 2000a, Hayward et al., 2001.

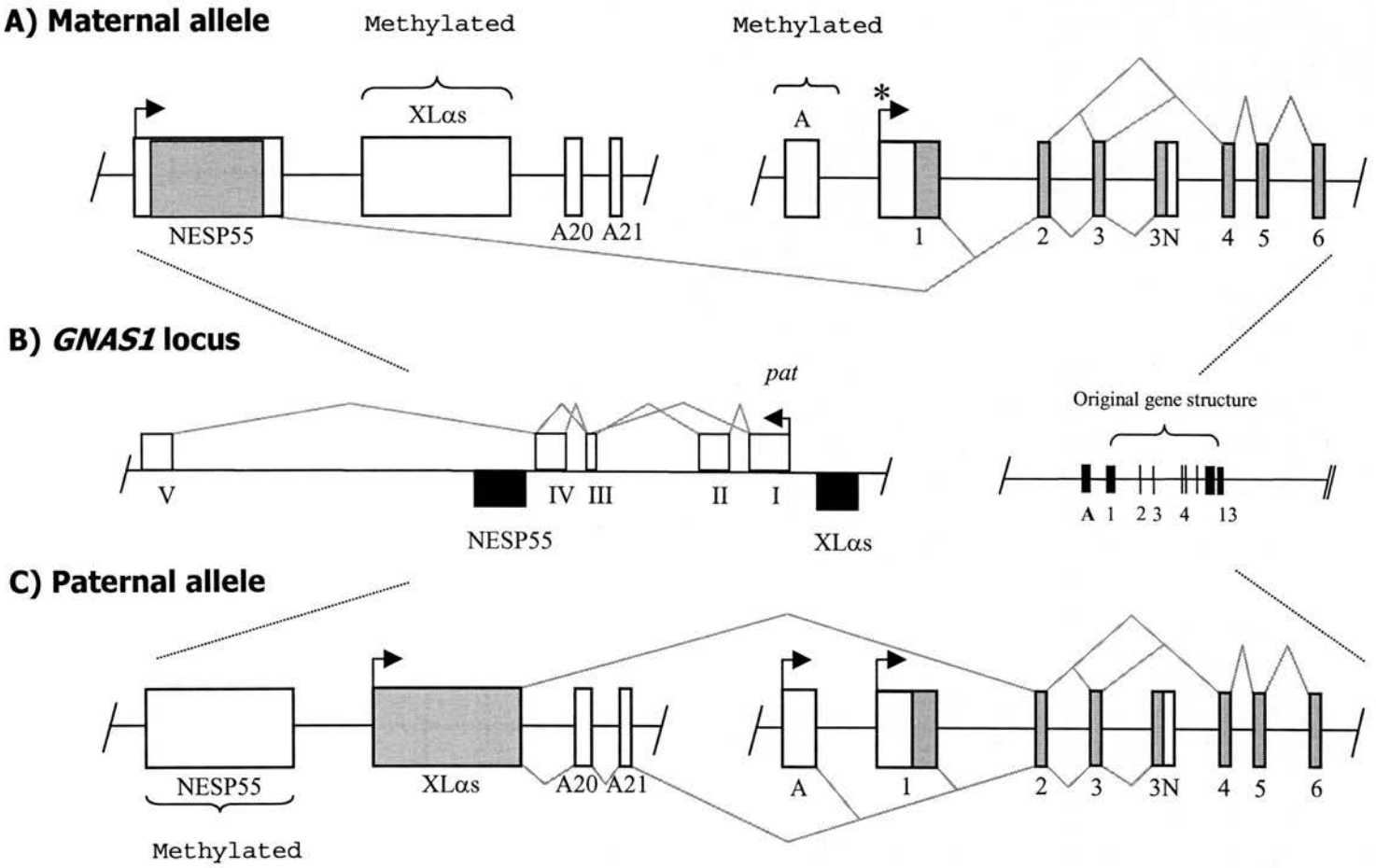


FIGURE 3.24. Splicing patterns observed for the *GNAS1* transcriptional unit

CHAPTER 4

SCREENING FOR NEW GENES IN SUBCLONE LIBRARIES DERIVED FROM PAC DNA CLONES THAT MAP TO CHROMOSOME 20q13

CHAPTER 4. SCREENING FOR NEW GENES IN SUBCLONE LIBRARIES DERIVED FROM PAC DNA CLONES THAT MAP TO CHROMOSOME 20q13

4.1 INTRODUCTION

There is considerable evidence that one characteristic of chromosomal regions regulated by genomic imprinting is the presence of clusters of imprinted genes. A further intriguing aspect, that must be of central relevance to the mechanisms for regulating imprinting in such regions, is the presence in at least two of the best characterised of these clusters (chromosomes 11p15 and 15q11-q13) of non-imprinted genes interspersed with imprinted genes (Buiting et al., 1995, Lee MP et al., 1999b, Ohta et al., 1999a and b).

How these imprinted genes are co-ordinately regulated is not entirely clear at present, but nonetheless the available information does clearly indicate that the chromosomal distribution of imprinted genes is non-random. As a consequence, several imprinted genes have been identified purely by positional criteria. It remains to be explained, however, why other genes mapping within these same regions (and even, furthermore, located between a pair of two similarly imprinted genes) nevertheless escape imprinting (Yuan et al.,1996, Zubair et al., 1997, Lee MP et al., 1999b, Ohta et al.,1999a and b).

One hypothesis that is often invoked to explain the coordinate regulation of an imprinted gene cluster involves the action of a long distance control mechanism in these regions. Regardless of the precise control mechanisms, though, a direct implication of the existence of these co-regulated clusters is that once a new gene has been characterised as imprinted, the likelihood that there may be other imprinted genes in the same chromosomal region can be investigated (Caspary et al., 1998, Lee et al., 1999b, Bell and Felsenfeld, 2000, Hark et al., 2000). We pursued this line of investigation in the case of distal chromosome 20q13. After characterising *GNAS1* as a complex imprinted gene (Hayward et al., 1998a and b), there was considerable

interest in identifying any other genes that may be in this region, upstream and downstream of *GNAS1*, in order to assess whether they might form additional components of a larger co-regulated imprinted gene cluster.

The approach that was taken to identifying new genes in this region was first to identify overlapping PAC clones that contained *GNAS1* and the regions upstream and downstream of it, and then to construct plasmid subclone libraries derived from these PAC clones. The inserts of a representative number of these subclones were then sequenced at both ends and used to search EST databases using the BLAST algorithm. The sequences were also analysed using the HGMP NIX interface to a variety of programs for predicting structural features of DNA sequences. In the longer term, once a putative novel gene was identified, further cDNA and genomic structure analyses were pursued, in order to determine the gene structure and address its imprinting status.

Initially, it was known from the screening of the RCPI1 PAC library that four clones (dJ60o16, dJ63n2, dJ110p14 and dJ309f20) contained sequences from *GNAS1*. Only one of these (PAC dJ309f20) contained the complete gene sequence, while the other three clones overlapped the downstream region of *GNAS1* (corresponding to the T7-end of dJ309f20). Subclone libraries in the vector pUC18 were constructed, with inserts derived from digestion of PAC dJ309f20 DNA with the restriction enzymes *Bam*HI or *Bgl*II. In addition, the sequences at the T7 (downstream) and SP6 (upstream of *GNAS1*) ends of PAC dJ309f20 were determined. Based on these sequences, oligonucleotides were designed for use in further PCR screening of the RCPI1 PAC library, in order to identify other clones that overlapped dJ309f20. Following this, further sequence analysis of subclone libraries constructed from these clones was carried out.

4.2 MATERIAL AND METHODS

4.2.1 POLYMERASE CHAIN REACTION

All PCRs and RT-PCRs were carried out with one of six different buffer conditions, as described in Chapter 2, Materials and Methods.

4.2.2 OLIGONUCLEOTIDES

The following oligonucleotides were used for PCR, and sequencing:

NAME	SEQUENCE 5′-3′
SEQUENCE ANALYSIS OF M13F/R ENDS OF SUBCLONES	
119-1R/A1	TGG ATG GCA GCA GGC AAA G
120R1/114	GCT GTG AAC TGT CCA TTT
34F8/114	GGA TGG TAA CTT CGG GCA G
1476-a	GGA GCC TTC GTT CTG TGC
4FHdF-a	CTT AGC CAT CCT GAG AGT CC
D7BgF-a	GCC TTC TGT GGG TTC TGT AGT TT
COMPLETE SEQUENCE ANALYSIS OF *BamHI* SUBCLONE G11	
THF	AAC AGC CCT GCA TCT GCC GT
THF2	AGG GGA ACA GAA AAC TTA GGC
G11THF3B	TAA AAA GGA AGG GGA TAC AGA T
THF4	CAA AGC CCC ACC ACC GT
7F	TTG ATG TGA GCG CTG TGA ACA
G117F2	CTG GAA TCT TTG AGA CCA AG
7R	TGC CAA TCT AAA GCA AAG GTC
7R2	AGA AGG TTA GGG CAA ACA GGA
7R3	GGT AAT GCC TGA CAA AAA CA
7R4	TTG GGG TAA TTA ATA TGG AA
G117R5	AGC CAG CCA AAC CCC AGC
G117R6	CCC GAC AAA ATT CAA CAA
COMPLETE SEQUENCE ANALYSIS OF cDNA CLONE 1430	
1430R2	CAT CGT CTG GGC TGA ATA A
1430R3	TAA GAC CTG TGG CAT CCC C
1430R4	ACT TTG GGA GGC CGA GGC

1430F2	TAG CCC TCT CCT TGT GGC A
1430F3	TGG GGA CAC ACT GGT TGC CT
COMPLETE SEQUENCE ANALYSIS OF cDNA CLONE 088f21	
88f21R2	ATC CAA CAC CTT TCC CCA GC
88f21R3	GGG GCG TAT CTA GTG TTG AA
88f21F4	GGA AGC TGA CAG GGA CCA A
88f21F3	GCC CCA TCT CCA ATT CTA AT
88f21R5	TGC CTC CTT TTC GAC GAC TGA T
88f21F5	TTG GAA GCA GCA GGG ACC
COMPLETE SEQUENCE ANALYSIS OF CDNA CLONE NX11	
NX11-EST1	GTT GGC GGT CAT CAT CA
NX11-EST2	GAA AGT TCT CCC ATT GAG AGC AC
NX11-3	TCT CTC TGA TGC CAC AAC A
NX11-4	GCA AAT GCA GGC TGA TGG CGC
NX11-5	TGT TGT GGC ATC AGA GAG A
NX11-6	GCG CCA TCA GCC TGC ATT TGC
NX11-7	TCT CCT GGC CAT ATC TAG C
NX11-9	CAG CTG CTG GGC CAC TCT TTG
NX11-11	CCG TGT TCA GGG TCG TTT CT
NX11-12	TTC CAA TCA TGC CAT TCT T
NX11-13	CCT GGA TGT GGG CTC TGC A
NX11-14	AAG AGG GAG CAG GGA GGA TC
NX11-15	GTT GGC GGT CAT CAT CA
NX11-16	GGA ACA CGT AGG GAG CAG AG
NX11-17	GCA CAA ACA TGC AGC CAA GT

4.2.3 PULSED-FIELD GEL ELECTROPHORESIS (PFGE)

For PFGE analysis the PAC clones were digested with rare cutting restriction enzymes, and electrophoresis was performed in 1% agarose gel using a CHEF-DRII apparatus (BioRad). The voltage was set at 200 volts, with a temperature of 14°C for 16 hours. Thereafter, the separated fragments were transferred to a nylon membrane (Hybond N+, Amersham) by capillary blotting and then hybridised to ^{32}P-labelled probe as described in Chapter 2, Materials and Methods.

4.2.4 ANALYSIS OF SUBCLONE LIBRARIES FROM PAC CLONES dJ309f20, dJ96n2 AND dJ654c22

PAC DNA was purified and used for the construction of the subclone libraries. Sequencing of plasmid subclones from the libraries was performed using either a ^{33}P-labelled dideoxy terminator kit (Amersham USB) or dye terminator chemistry on an ABI377 automatic sequencer. The sequences obtained were used for database searching with the BLAST algorithm, and further analysis using the HGMP NIX interface to multiple sequence analysis programs, as described in Chapter 2, Materials and Methods.

For screening of the subclone libraries by hybridization, the subclones were replicated from 96-well plates onto nylon filters. These filters were hybridized using the Quick oligo hybridisation method, as described in Chapter 2, Materials and Methods.

4.3. RESULTS

Four PAC clones (dJ60o16, dJ63n2, dJ110p14, dJ309f20) had originally been identified in the RPCI1 library by PCR screening for exon 7 of *GNAS1*. However, further analysis indicated that of these four clones, dJ309f20 was the only one that contained all the exons of *GNAS1*. It was therefore chosen for further sequence analysis, and as the starting point for extending the PAC contig.

4.3.1 SCREENING OF PAC dJ309f20 SUBCLONE LIBRARIES

A subclone library in the pUC18 vector was constructed from PAC dJ309f20, digested with the restriction enzymes *Bam*HI or *Bgl*II. 50 subclones from each of the *Bam*HI and *Bgl*II libraries were picked at random and both ends were sequenced using the oligonucleotide primers M13F and M13R. The sequences obtained were used to search (using the BLAST algorithm at NCBI) four different databases: non-redundant GenBank, human ESTs, mouse ESTs and other (non-human or mouse) ESTs. From these initial 100 subclones analysed the results were as follows:

Seven subclones contained insert sequences identical to various exons of the *GNAS1* gene; four contained ESTs other than *GNAS1*; 21 had sequences similar to *E. coli* (presumably as a result of the presence of *E. coli* genomic DNA contaminating the preparations of PAC DNA); one had a sequence similar to *S. cerevisiae*; ten insert sequences were not of good quality for analysis and the remaining 57 subclones identified sequences without matching EST or related to repetitive elements type L1 or *Alu* (Table 4.1).

NAME	***GNAS1***	**ESTs**	***E. coli***	***S. cerevisiae***	**Rep. elements**	**Other**
NUMBER	7	3	21	1	57	11

Table 4.1. dJ309f20 subclone library screening results.

The three novel ESTs identified in this initial round of sequencing were in the inserts of subclones A1, G11 and C6 in the *Bam*HI library and these subclones were therefore further analysed.

4.3.2 ANALYSIS OF SUBCLONE BamA1

The insert in subclone A1*Bam*HI group 1 (referred to hereafter as BamA1) was sequenced at both ends. The sequence obtained from the M13R end was 300 nt in length and contained no EST matches. The sequence obtained from its M13F end was 520 nt and had a significant match to the EST AA159229, representing an mRNA derived from human pancreas.

The IMAGE cDNA clone corresponding to EST AA159229 was obtained from the UK HGMP Resource Centre under its identifier 1430-d08, (referred to hereafter as cDNA clone 1430). This cDNA had been cloned into the pBluescript SK vector using the enzymes *Eco*RI and *Xho*I and was therefore digested with both enzymes, showing that its insert was ~1kb in length (Figure 4.1). DNA from the genomic subclone BamA1 was digested with the *Bam*HI, showing that its insert length was ~12kb.

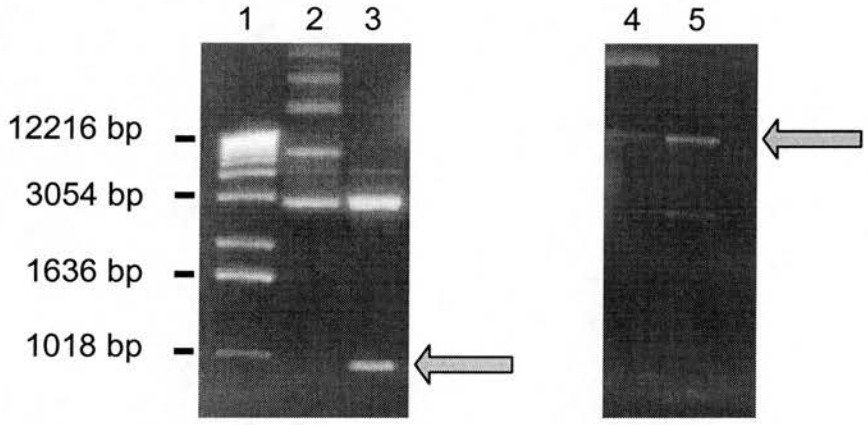


Figure 4.1. Restriction digests of cDNA clone 1430 and genomic subclone BamA1. Lanes: 1) 1 kb ladder, 2) Clone 1430 uncut, 3) Clone 1430, *Eco*RI+*Xho*I digested, 4) Subclone BamA1 uncut, 5) Subclone BamA1, *Bam*HI digested. The arrows mark the insert bands.

The insert of cDNA clone 1430 was then completely sequenced on both strands. Its exact length proved to be 964 nt (Figure 4.2).

(M13R) *Eco*RI

1 ***GAATTC***GGCACGAGAGAGGTTACCAACTTCACAAATAATGGGTTTCCATAGAGGATGTCATCTTGAAAAA

71 CACAAACATAGTTTCATGTGTTTCATTGGGCTTCACATTGCAGCCTTTTCATCTATCTACATATCTTGGA

1430R2---------

141 GACCCTTTCCTGCTGATTCCTGTAGGATTTGCCTTGTTCTATTTCACCATCACCTCATCGTCTGGGCTGA

--->

211 ATAAAGACAGTCAGCAGTCACATCCCTGTCGACGTGCATTCAGCTCTTGCCACTTTGTCCCTGTTAGAAA

281 CTGCACTGCTGACCATGTCGTAGTACACACCTCCATGTAAATGTCACAAGTGTTTCTCTGGAGATGATAT

<---------

351 CAAGCAGTGGAGCTGCTCATAGGCCAACACAGAGCCTTGGCTACACTGCACTTCTCTACGAGGCAACCAG

----1430F3

421 TGTGTCCCCACCAGCAGGGCCTGAGTATTTGTCTCCCATCCCCTGCTAGCACACGACCCTCTTCTTTCTG

1430R3------------>

491 GTAGAGGAGACAGTAAGACCTGTGGCATCCCCAGGCCCTCTATGAGTCAGGAGAGAGGGCATGGAGCTGT

561 TTTCCTGCTCTGCTTGTGGGGCTGGGTTTCTGAGGCCCATGACCTTGGCTTCATCAATATTTTCTCTAGC

<------------1430F2

631 CTGTCAAGAAAAGTTGAACTTCAGTGAAATATATCTCAGAAATTTGCCACAAGGAGAGGGCTAGGGAAAA

701 CCTAACTAAAAGATGTTAGAAAAGAGAGGCCAAGTGTATCCCAGGTTCTGTGTCCAGTTTCTTTTAGTCC

771 TTCATGAAGGTGACTGAGCTTGTTTTAGAAATAACATAGTTCAGCCAGGTGCGGTGCCTCACACCTGTAA

1430R4------------>

841 TCCCAGCACTTTGGGAGGCCGAGGCAGGCAGATCACTTGAGGTCAGGAGTTCAAGACCAGCCTGGCAACA

911 TGATGAAACACCATCTCTACTAAAAATACAAAAAAAAAAAAAAAAAAA***CTCGAG*** *Xho*I (M13F)

Figure 4.2. Complete sequence of cDNA clone 1430. The *Eco*RI and *Xho*I sites (corresponding to the M13R and M13F end of the subclone) are marked in bold italic letters. The position and orientation of the oligonucleotides used for sequencing are indicated by the arrows and are underlined.

The genomic insert in BamA1 was also further sequenced from its M13F end and the sequence obtained was compared with the complete sequence of cDNA clone 1430. It was found that the sequences were oppositely orientated at their M13 ends and that a probable intron-exon junction was present. The alignment of these sequences is illustrated in Figures 4.3 (apparent 5′ end of the cDNA clone) and 4.4 (apparent 3′ end).

```
                                  Intron                          Exon
1) BamA1  M13F-end    ..TTCATACATCACCTCTCAAATCCCCC**AG** | AGGTTACCAACTTC
2) 1430   M13R-end      CCCGGGCTGCAG**GAATTC**GGCACGAGAG | AGGTTACCAACTTC
                                  Vector                          cDNA
```

Figure 4.3. Comparison of the M13R end of cDNA 1430 with the matching part of genomic subclone BamA1. The *Eco*RI cloning site of the cDNA is in bold letters and underlined and the adaptor sequence using for cloning is underlined. The intron-exon junction in the genomic sequence is marked by a vertical line and is in bold letters.

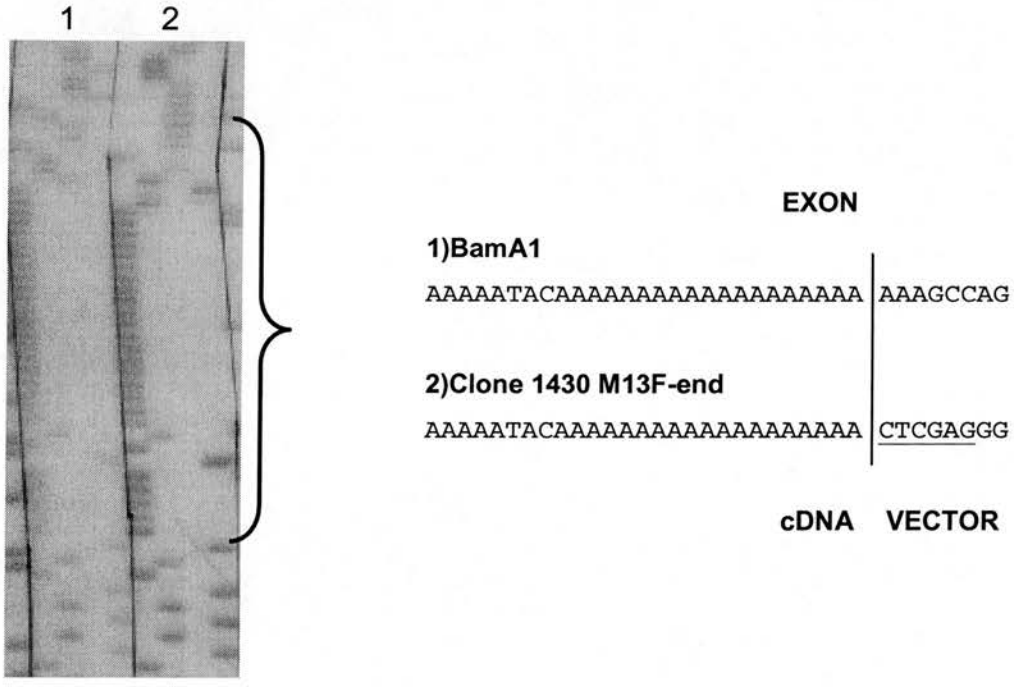


Figure 4.4. Autoradiograph showing partial sequences of a region from the genomic subclone BamA1 (sequence 1) with the matching part of the EST clone 1430 M13F end (sequence 2) derived using the primer 1430R4. The partial sequences indicated by the brace are compared on the right of the figure. The site where the sequences diverge is marked by a vertical line and the *Xho*I cloning site of the cDNA clone is underlined.

Further genomic sequence analysis of PAC dJ309f20 was later carried out by Dr. B. E. Hayward, including determination of the complete sequence of the 12 kb genomic insert in subclone BamA1 (Hayward and Bonthron, 2000). This data indicated that the putative transcript represented by cDNA 1430 lies within a region ~19kb upstream of *NESP55* (discussed in Chapter 3).

4.3.3 ANALYSIS OF SUBCLONE BamG11

Sequencing of the insert of dJ309f20 subclone G11*Bam*HI (referred to hereafter as BamG11) identified ESTs at both M13-ends. The sequence obtained from its M13R end was 271 nt in length and included part of exon 9 and the whole of exon 8 of *GNAS1*. The sequence obtained from the M13F end was 352 nucleotides in length and had 3 significant EST database alignments: AA195902, an mRNA derived from muscle, H72932, an mRNA from fetal liver and spleen, and D81936, a mRNA from fetal brain. The sequence of EST AA195902 was used for a further BLAST search, identifying another three ESTs, AA284401, AA278187 and AA253299.

These EST sequences were also used to search the TIGR database, identifying the 1255 nt consensus sequence THC205478 as a significant match, with a percentage of identity of 97.8%. Based on the existence of this apparent EST contig, it was decided to characterise in more detail the insert of genomic subclone BamG11. BamG11 DNA was digested with *Bam*HI, showing that its insert was ~2kb in length (not shown).

From the known genomic structure of *GNAS1*, the position of the *Bam*HI site in exon 9 made it clear that the 2 kb insert in BamG11 must also contain exons 7 and 8 of the gene. To establish the distance separating the sequence in BamG11 that matches THC205478 from exon 7 of *GNAS1*, a PCR was performed with primers named THC (THC sequence) and PHP7R (exon 7 of *GNAS1*). As template, BamG11 DNA, PAC dJ309f20 DNA and human genomic DNA were used. The PCR product obtained in each case was ~1200 bp (Figure 4.5).

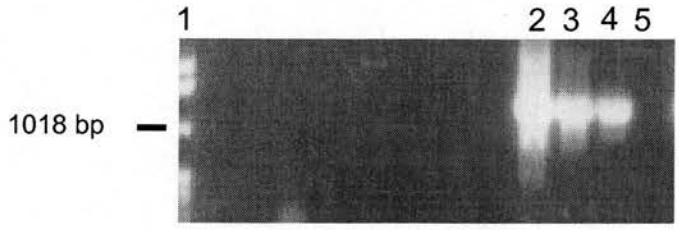


Figure 4.5. THC-PHP7R PCR products. Lanes 1) 1 kb ladder, 2) BamG11 DNA template, 3) dJ309f20 PAC DNA template, 4) human genomic DNA template and 5) no template (control blank).

Finally, the BamG11 insert was completely sequenced on both strands. The sequence was 2103 nt in length (Figure 4.6).

Figure 4.6. Complete sequence of genomic subclone BamG11. The *Bam*HI sites are underlined. The sequence contained in THC313290 (the most recent version of the original cDNA contig THC205478) is in bold letters and *GNAS1* exons 7 to 9 are highlighted in grey and labelled. The arrows mark the positions and orientations of the oligonucleotides used for analysis.

M13F

1 GGATCCAAGTCAGAAGTTAAGTACAGGTACAAGATTAAGAGAGTTTAACCACCCCAGTTTTGAATGCTGG

<----------G117R6

71 GGGGGCAGGGAGACCCAGTTTCGTTAATTAACAGTAGCTTAGCCAGATTGTTGAATTTTGTCGGGTTTCG

141 TTTTCTCTCTCAAATCATTTAGAAGTTTTTGGGGTTTTTTTAAGCAACACTTAATTACTCCTGAAACTTT

211 GTCTGAAAACGCACCATTTGTATAGATCATGAAAAGTTTTAAGGAAACTCAGAGAAAAAGAGAACAACGC

THF---------->

281 AGCTTAAAACTTTTAAAATGTCCTCCCTCACCCGTGGCTCAAACAGCCCTGCATCTGCCGTGGCCGGCAC

351 GTTTCTGGTTGAACTGCCTTTATGTTAAAGTTCAGATACTGGTAGTGTGCCCATTTCTTAAGCTGTCTAT

<----------G117R5

421 TTTTATTTGTTGAGCTGGGGTTTGGCTGGCTCCACTCCAGATGTCTCTCTCACAAGATTTGGTGCTGATG

491 ATCTATTTATAGAACTGTGGTTCTGTTGCCATGGTAACATGCTGGAGGCCAGGGCGGCTGGGGAGCTATT

561 TCTGGACTGGTGCTGTAATGTAAGATTGATTGGGCAAGTTAGTATATCCTCTAAGCCAGACTAACTCTGA

THF2---------->

631 ACTAGTAAAAAGGAAGAGGGGAACAGAAAACTTAGGCAGTTTCTTTAAATAAACTTTTCTCTCTTTATGA

701 TTTTCTTTTCTCGTTAGCCCGCTTTAAAACAATTCCAATCTCTACATGCCCCTCCCTCCAAAAAATAACT

<----------7R4

771 GGTTTTAACATTAATTTTCCATATTAATTACCCCAATCTTTCAAAAGTAAATTTTCCTGTGTGTCTAGTC

841 AAGCAACACAAACAAGATGCTTTTTTTAATGAAAAGCGTAATATCTGGAGTGTTCTATTTCATGGACCAA

911 ACAGAACGGAAGAGAACTTTGTGTTTGTTTCCTGTTCAGGATGGCTAGGGCGAGAAGGGCCCCTTTGTGC

G11THF3b---------->

981 CAACCTCTTTTGTTCTCTTTAATCAATGAGATTCTTAAAAAGTAAAAAGGAAGGGATACAGATTCTCAGT

1051 AACTAAAACAATCTCGTGTGCCCTTGAGGGGAAAGTCCTTGATGTTTTTAAGAATGTCACTTTATTGTTT

1121 TTTAACTGAATGATATAGAGGTATACAATTTTCAAACTGTTTGCCATTTTAATCAAGCAATTTGAAAATT

<----------7R3

1191 AAAATGTTTTTGTCAGGCATTACCAAATGGCACAGAATGTGATAGGCCAGCCTGGTTTTGGGGTCCTTTC

1261 TCTACTGGTTGATATGCATAAAACCTTCTAAAAATCAAGAATATTGCCAGAGAGCAACAGGAATAAAGAA

THF4----------

1331 **GCTAAGTAAAGAATAAAAAAGAAAAATAG**AAAAAATAAAAATAAACACGAAGAACAAAGCCCCACCACCG

>

1401 TGCTGTGCTGTTTGTGTGGCCCCACTGCGTCGAGGCCACAGGCTAGCTGCTAGACGCATCTAGAGTTCCC

<----------7R2

1471 TGATTCCTAAAATTATTTATCTTAAATCCTGTTTGCCCTAACCTTCTTAAGGCATCAGCTTTGAGTTACA

1541 AATGTAACCAACACACAAGCAAATGTGCCATTGACTTAGTGCTGCATAACTGTGGGACGGTCACTTCCGT

7F---------->

1611 TGAGCCTGACCTTGTAGAGAGACACAAATAGTTGGCAAATTGATGTGAGCGCTGTGAACACCCCACGTGT

1681 CTTTCTTTTTCTCCCAGCTTCCTGGACAAGATCGACGTGATCAAGCAGGCTGACTATGTGCCGAGCGATC

GNAS1 EXON7

<----------

1751 AGGTGTGCAAAACCCCTCCCCACCAGAGGACTCTGAGCCCTCTTTCCAAACTACTCCAGACCTTTGCTTT

-------7R

1821 AGATTGGCAATTATTACTGTTTCGGTTGGCTTTGGTGAGATCCATTGACCTCAATTTTGTTTCAGGACCT

GNAS1

G117F2---------->

1891 GCTTCGCTGCCGTGTCCTGACTTCTGGAATCTTTGAGACCAAGTTCCAGGTGGACAAAGTCAACTTCCAG

EXON 8

1961 TAAGCCAACTGTTACCTTTTTATATAACAGAGATCATGGTTTCTTGACATTCACCCCAGTCCCTCTGGAA

2031 TAACCAGCTGTCCTCCTCCCCACCAGCATGTTTGACGTGGGTGGCCAGCGCGATGAACGCCGCAAGTGGA

GNAS1 EXON 9

2101 TCC*M13R*

This completed sequence was used to perform further database searches using BLAST, showing that it contained the entire sequence of putative cDNA contig THC205478, as well as *GNAS1* sequences corresponding to part of intron 6, exon 7, intron 7, exon 8, intron 8 and part of exon 9 (Figure 4.7).

Figure 4.7. Subclone BamG11. A) *GNAS1* partial gene structure (not to scale), as in Kozasa et al. (1988). Exons are represented by grey boxes and are numbered. The *Bam*HI sites are indicated. B) Genomic subclone BamG11, showing the position of the THC313290 contig sequence within intron 6 of *GNAS1* and the *GNAS1* exons contained within the subclone. Map to scale.

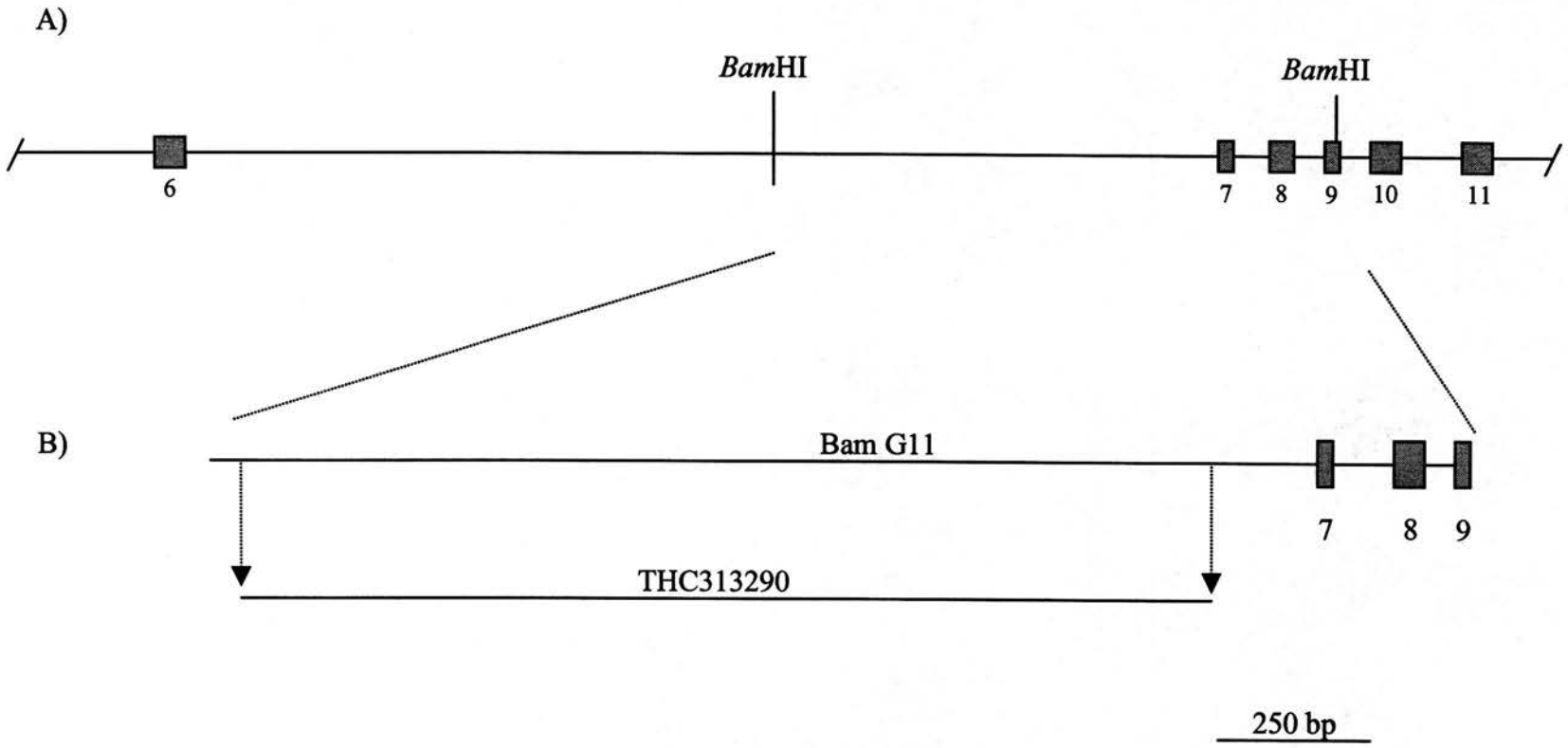


Figure 4.7. Subclone BamG11

More recently, the THC205478 entry has been replaced by an updated contig, THC313290. The sequences of genomic subclone BamG11 and THC313290 were compared using BLAST 2. The sequences were 99% identical, with minor differences between them probably representing sequencing errors. As shown in figure 4.6, (a) there is a single contiguous region of match between the genomic sequence and THC313290 (b) the genomic sequence has a poly(A) stretch corresponding in position to that of the putative EST contig. The results suggest that THC313290 actually represents a genomic sequence, represented in the EST databases as a result of priming from a (dA)-rich region of DNA. However, it is also possible that the THC313290 ESTs represent a genuine unspliced transcript, which could also be part of a larger transcript. The 3′ end of the sequence THC313290 is 336 nt away from exon 7 of *GNAS1*.

4.3.4 ANALYSIS OF SUBCLONE BamC6

The sequence obtained from the insert in subclone C6 *Bam*HI (referred to hereafter as BamC6) was 158 nt in total length. This sequence showed a significant match to a human EST T84962/11764, derived from a fetal liver and spleen cDNA library.

4.3.4.1 Clone 088f21 enzymatic digestion analysis

The corresponding IMAGE cDNA clone was obtained from the UK HGMP Resource Centre under its identifier 088f21. The cDNA insert was cloned between *Pac*I and *Eco*RI sites in the vector pT7T3D. Plasmid DNA was extracted and digested with *Pac*I and *Eco*RI, yielding a band of ~700 bp .

4.3.4.2 Sequence analysis of clone 088f21

The insert of clone 088f21 was then completely sequenced on both strands (Figure 4.8).

Figure 4.8. Sequence of the insert of cDNA clone 088f21. The *Bam*HI, *Pac*I, *Eco*RI and the poly(A) signal sites are marked in bold. The sequence contained in subclone BamC6 is highlighted in grey and the position and orientation of the oligonucleotides used for the analysis are marked by arrows. The splice junction (described in section 4.3.5) is marked by a vertical line.

When the sequence from clone 088f21 and subclone BamC6 inserts were compared, it was noticed that they had the same orientation with respect to their M13-ends and also that at least one intron-exon junction was present (Figure 4.9).

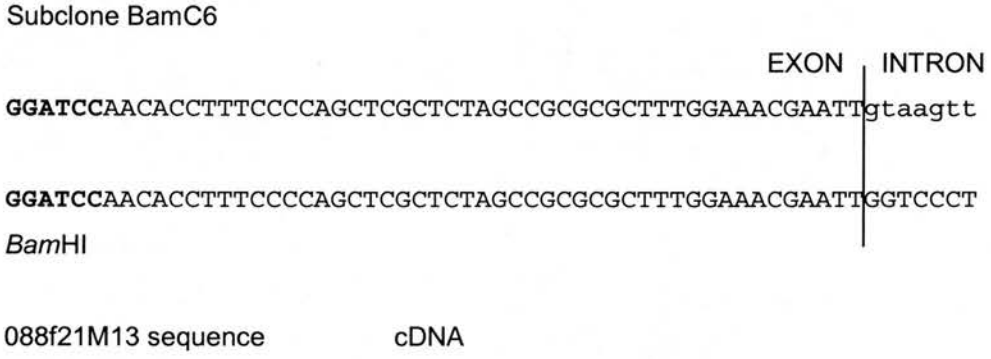


Figure 4.9. Alignment of partial sequences from cDNA clone 088f21 and genomic subclone BamC6. The intron-exon splice junction is marked by a vertical line, and the intron sequence in subclone BamC6 is in lower case. The *Bam*HI site is in bold.

A further hybridisation screen of the dJ309f20 genomic subclone libraries was performed to identify subclones containing the remaining part of the sequence of cDNA clone 088f21. For this purpose, a PCR product was generated using oligonucleotides 88f21F3 and M13R, purified from a 1% low-melting-point agarose gel, and labelled as described in Chapter 2, Materials and Methods for use as a hybridisation probe.

Several subclones were identified in this screen (Figure 4.10). Among them, as expected, were BamC6 (positive control) and, unexpectedly, the subclones BamA1 and Bgl34. The subclone BamA1 was mentioned above because it contained sequences corresponding to the cDNA 1430 (section 4.3.2) and was sequenced completely later in our laboratory. The subclone Bgl34 was known to contain the *NESP55* sequence (Chapter 3) and its sequence was also completed in totality later.

The oligonucleotide 88f21F3 itself was also used for a further round of screening of the subclone libraries. Since it corresponds to a sequence adjacent to the poly(A) signal in cDNA clone 088f21, the genomic subclones identified by it should contain the most 3′ region of this transcript. In the *Bgl*II library clones F9, D2 and C2 were positive.

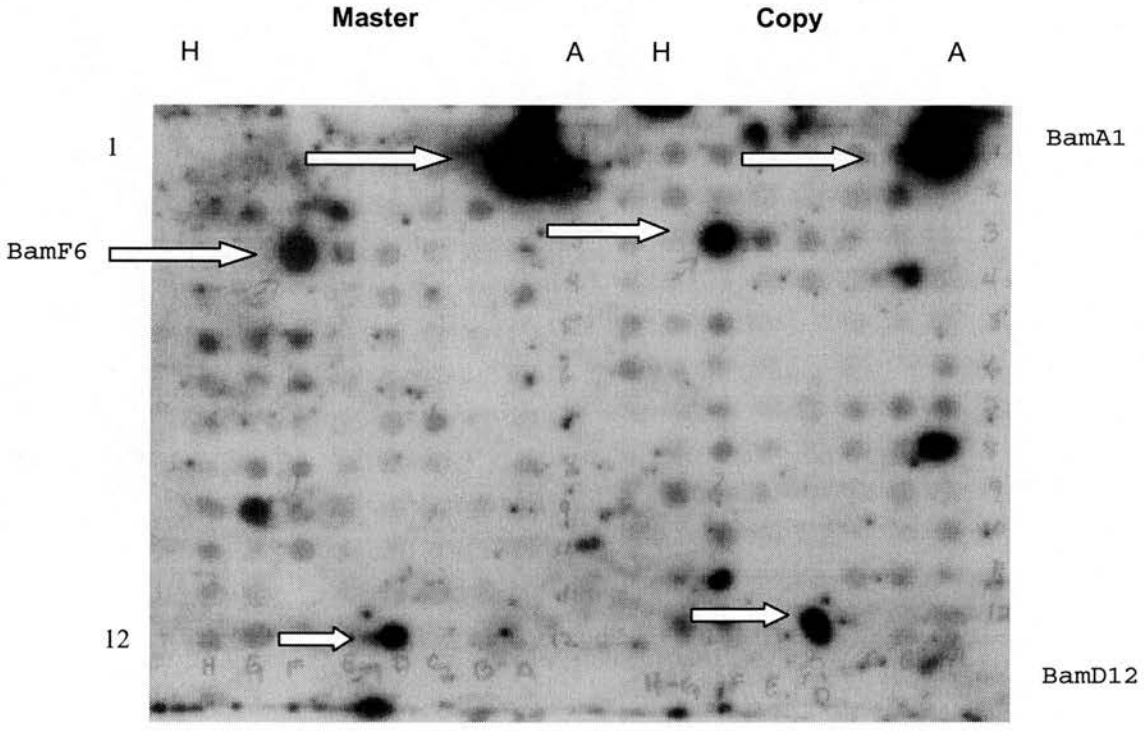


Figure 4.10. dJ309f20 subclone library screening for clones containing exons of the antisense transcript 088f21. The M13R-088f21F3 PCR product was used as probe. Duplicate membranes were hybridized; both master and copy membranes are shown. The positive subclones, BamA1, BamD12 and BamF6 (in this group) are marked by block arrows.

4.3.5 Analysis of subclone BglC2

From the several subclones positive for the 3′ end of the antisense transcript 88f21, subclone C2 from the *Bgl*II library (referred to hereafter as subclone BglC2) was further analysed. By PCR with the M13F and M13R primers an insert length of ~2.7 kb was determined.

Primer 88f21F3 (see Figure 4.8) was used for sequencing on this new genomic subclone. When the 301 nt genomic sequence obtained was compared to the 088f21 cDNA sequence the two sequences were identical as far as the already identified intron-exon splice junction (Figure 4.11; sequences shown on reverse strand to correspond to Figure 4.8). Immediately beyond this point the sequences diverge, with a consensus ---AG/ splice acceptor site being present in the BglC2 sequence.

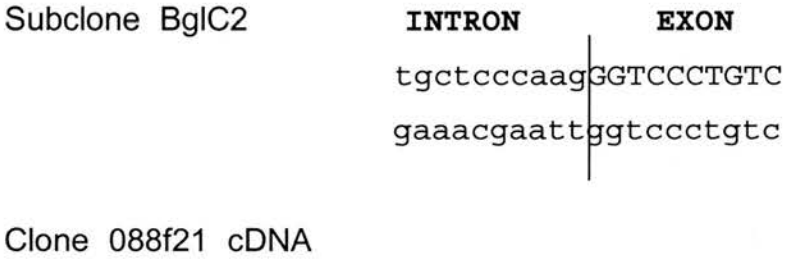


Figure 4.11 Comparison of part of the sequences of genomic subclone BglC2 and cDNA clone 088f21. The intron-exon splice junction is marked by a vertical line.

As mentioned, the large (12 kb) subclone BamA1 was also positive on screening with oligonucleotide 88f21F3, and from its size, it was expected that it would contain the complete 2.7 kb *Bgl*II insert of subclone BglC2. In order to establish the position of the 088f21 cDNA sequence (later identified as exon V of the antisense transcript at the *GNAS1* locus, Hayward and Bonthron, 2000) within these two genomic subclones and also their respective orientations relative to the M13 primers, PCRs were set up using the M13F or M13R primers with 88f21F3 or 88f21R3. The deduced map of these genomic clones is shown in Figure 4.12.

Figure 4.12. Diagram to show relative positions and orientation of subclones BamA1 and BglC2. The orientation of the vector relative to the subclone inserts ends is indicated by M13F and M13R. The *Bam*HI and *Bgl*II sites are shown (in subclone BglC2 the *Bgl*II sites are shown in brackets, as they were destroyed during cloning into the *Bam*HI site of pUC18). The oligonucleotides used for sequencing and PCR analysis and their orientation are indicated as arrows. Exon V of the antisense transcript is indicated by the grey box. Diagrams are to scale.

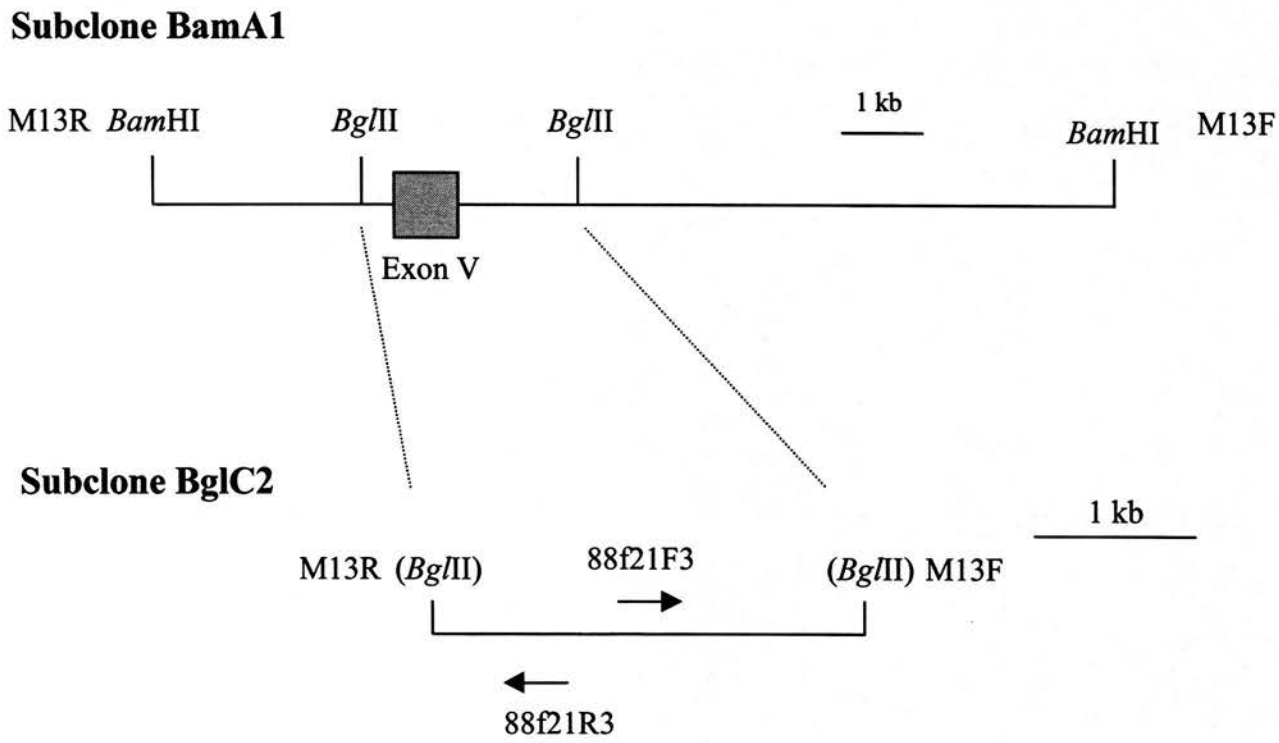


Figure 4.12. Diagram to show relative positions and orientation of subclones BamA1 and BglC2

4.3.6 ANALYSIS OF SUBCLONE BAMF3

The original sequence of genomic subclone BamC6 (which had identified the 088f21 cDNA in the EST databases) consisted only of a small insert of 158 bp between two *Bam*HI sites. Subsequent screens for other genomic clones containing parts of cDNA 088f21 identified among others subclone Bg134. This subclone had already been analysed, and was known to contain the NESP55 exon (described in Chapter 3). However, restriction mapping and then later complete sequencing of Bgl34 revealed only one *Bam*HI site corresponding to the site at one end of BamC6. The position of this site indicated a discrepancy between the structures of BamC6 and Bg134.

To resolve this discrepancy in the presence or absence of the second *Bam*HI site between the two clones, another *Bam*HI subclone that was positive on screening for 088f21 was analysed. This was subclone F3 (referred to hereafter as subclone BamF3).

Digestion of subclone BamF3 with *Bam*HI showed that unlike BamC6, its insert was ~7kb (data not shown). It was sequenced at both ends using M13F and M13R primers. The sequence obtained from the M13R end of this clone was identical to that of BamC6 as far as the position of the discordant second *Bam*HI site. At this point, no *Bam*HI site was present in subclone BamF3, consistent with the data obtained from sequencing of subclone Bgl34 (Figure 4.13).

It therefore appeared that the small insert ofsubclone BamC6 was artefactually generated as a result of an unusual aberrant fragmentation and cloning of a truncated *Bam*HI fragment.

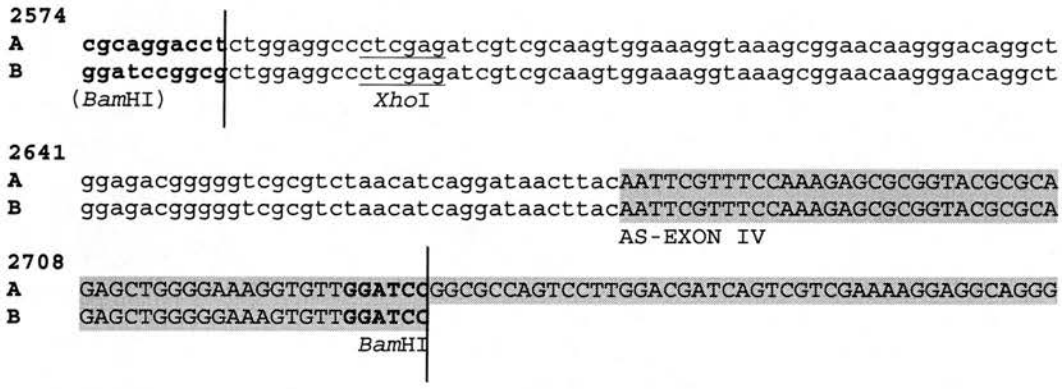


Figure 4.13. Sequence comparison of subclones BamC6 and Bgl34. The partial sequence (sequence A) of subclone Bgl34 is shown as in Figure 3.17, from nt 2574 to 2774, and it is compared to the complete sequence of subclone BamC6 (sequence B). The sequences are identical from nt 2584 to nt 2731 (region delimited by vertical lines, including the *Bam*HI-restriction site marked in bold letters and upper case-present in both subclones). This sequence comprises part of an intron (lower case letters) and part of exon IV of the *GNAS1* antisense transcript (marked in upper case letters and in grey). The sequences differ from nt 2574 to 2573 (marked in bold letters), where an artefactual *Bam*HI site is present in BamC6 (shown in brackets). This result was also verified by sequencing on subclone BamF3.

4.3.6.1 Further sequence analysis of subclone BamF3

A 333 nt sequence read was obtained from the M13F end of subclone BamF3. This sequence yielded two significant alignments in the EST databases, to accession numbers AA179082 and AA176863, whose sequences overlap. The IMAGE clone corresponding to EST AA179082 was obtained from the UK HGMP Resource Centre under its identifier 1476-k11.

4.3.6.1.1 Analysis of IMAGE clone 1476-k11

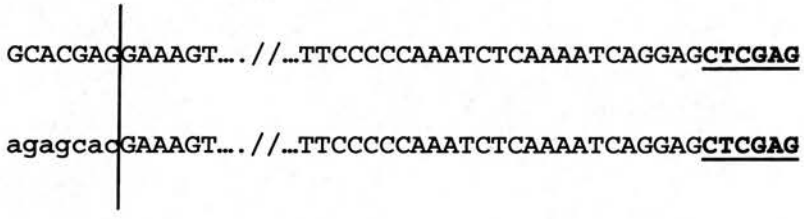
The cDNA insert of 1476-k11 had been cloned into a pBluescript vector with the enzymes *Eco*RI and *Xho*I. An *Eco*RI+*Xho*I digest of this clone showed that the insert size was only ~250 bp. The insert in clone 1476-k11 was completely sequenced from both ends; it was 236 nt in length (Figure 4.14).

GAATTCGGCACGAGGAAAGTAAGGGAGCTCCTGGGAAGACAGCTCCGAGCAATCTCTACGCACAGAAC
GAAGGCTCCTTTCTGATCCTACGCCCTTTAAAAACGAATGGCACAAATACAGTATTGTAATCCCAAAA
ATGGCCTGGACAGAAACCAGCCATGTCAAAAAGGTGATTATTTCCCTCTATTGTTTGCTTTTAGCTCT
AGATGGCATTCCCCCAAATCTCAAAATCAGGAG**CTCGAG**

Figure 4.14. Clone 1476-k11 insert sequence. The *Eco*RI and *Xho*I sites are underlined and in bold.

The insert sequences from clone 1476-k11 and subclone BamF3 were then compared. As can be noted in Figure 4.15, clone 1476-k11 is included almost in its totality in subclone BamF3 and it may be part of a bigger transcript. The putative cDNA sequence of clone 1476-k11 was completely colinear with the genomic sequence of subclone BamF3 apart from its first 7 nt (Figure 4.15). These 7 nt are in fact derived from a linker used for library construction. In addition, the *Xho*I site at the other end of the cDNA clone appears not to be linker derived, but to correspond to an *Xho*I site present in the genomic sequence at this point. It is therefore possible that 1476-k11 does not represent a true cDNA species.

1476-k11 cDNA **VECTOR**



Subclone BamF3

Figure 4.15. Sequence alignment of EST clone 1476-k11 and subclone BamF3. The *Xho*I site is in bold and underlined. The site where the sequences diverge is marked by a vertical line.